BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 4: e50889
doi: 10.3897/biss.4.50889

OPEN ACCESS

Standards

Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data

Arthur D Chapman[‡], Lee Belbin[§], Paula F Zermoglio[|], John Wieczorek[¶], Paul J Morris[#], Miles Nicholls[¤], Emily Rose Rees[¤], Allan Koch Veiga[«], Alexander Thompson[»], Antonio Mauro Saraiva[«], Shelley A James[^], Christian Gendreau[˅], Abigail Benson[¦], Dmitry Schigel[˅]

‡ Australian Biodiversity Information Services, Ballan, Australia
§ The Atlas of Living Australia, Carlton, Australia
| VertNet, Buenos Aires, Argentina
¶ Museum of Vertebrate Zoology, University of California, Berkeley, United States of America
# Museum of Comparative Zoology, Harvard University, Cambridge, MA, United States of America
¤ Atlas of Living Australia, Canberra, Australia
« University of Sao Paulo, Sao Paulo, Brazil
» iDigBio, Gainesville, United States of America
^ Department of Biodiversity, Conservation and Attractions, Western Australian Herbarium, Kensington, WA, Australia
˅ Global Biodiversity Information Facility - Secretariat, Copenhagen Ø, Denmark
¦ U.S. Geological Survey, Lakewood, CO, United States of America

Corresponding author: Arthur D Chapman (biodiv_2@achapman.org)

Academic editor: Gail Kampmeier

Received: 07 Feb 2020 | Accepted: 16 Mar 2020 | Published: 20 Mar 2020

Citation: Chapman AD, Belbin L, Zermoglio PF, Wieczorek J, Morris PJ, Nicholls M, Rees ER, Veiga AK, Thompson A, Saraiva AM, James SA, Gendreau C, Benson A, Schigel D (2020) Developing Standards for Improved Data Quality and for Selecting Fit for Use Biodiversity Data. Biodiversity Information Science and Standards 4: e50889. https://doi.org/10.3897/biss.4.50889

Abstract

The quality of biodiversity data publicly accessible via aggregators such as GBIF (Global Biodiversity Information Facility), the ALA (Atlas of Living Australia), iDigBio (Integrated Digitized Biocollections), and OBIS (Ocean Biogeographic Information System) is often questioned, especially by the research community.

The Data Quality Interest Group, established by Biodiversity Information Standards (TDWG) and GBIF, has been engaged in four main activities: developing a framework for the assessment and management of data quality using a fitness for use approach; defining a core set of standardised tests and associated assertions based on Darwin Core terms; gathering and classifying user stories to form contextual-themed use cases, such as

© Chapman A et al. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

species distribution modelling, agrobiodiversity, and invasive species; and developing a standardised format for building and managing controlled vocabularies of values.

Using the developed framework, data quality profiles have been built from use cases to represent user needs. Quality assertions can then be used to filter data suitable for a purpose. The assertions can also be used to provide feedback to data providers and custodians to assist in improving data quality at the source. A case study, using two different implementations of tests and assertions based around the Darwin Core "Event Date" terms, were also tested against GBIF data, to demonstrate that the tests are implementation agnostic, can be run on large aggregated datasets, and can make biodiversity data more fit for typical research uses.

Keywords

data quality, profile, framework, fitness for use, standards, tests and assertions, data quality tests, vocabularies, Darwin Core, GBIF

1. Introduction

Biodiversity Information Standards (TDWG) is a not-for-profit volunteer-based scientific association formed to establish international collaboration among the world's biological databases (TDWG 2007). TDWG encourages the wider and more effective dissemination of information about biological organisms for the benefit of the world at large through the establishment of biodiversity information standards. In recent years, TDWG has focused on the development of standards for the exchange and dissemination of different types of biological and biodiversity data—including names, taxa, specimens, observations, images, geographic locations, ecology, genetics, traits, and animal movements.

The Global Biodiversity Information Facility (GBIF) is an international network and research infrastructure that aggregates biodiversity data shared by myriad sources around the world. The volume of aggregated biodiversity data has increased in recent years, with GBIF now publishingover 1.3 billion records (GBIF 2018, GBIF 2020). Quality varies considerably within this mass of data (Gaiji et al. 2013, Mesibov 2013, Mesibov 2018) and issues and variation in quality affect the fitness for use of these data in different contexts (Chrisman 1991, Chapman 2005a, Chapman 2005b).

Recognising the urgent need to address the data quality issue, TDWG, in conjunction with GBIF, established a Data Quality Interest Group to examine biodiversity data quality and to make recommendations on ways to address it (Belbin et al. 2013, Saraiva and Chapman 2013).

2. Background

Digital exchange of institutional biological data began in the 1970s (Busby 1979) with small amounts of data, largely between individual institutions and researchers. It wasn't until the 1990s that biodiversity data began to be digitised on a large scale and made available to a wider audience (e.g., ERIN (Chapman and Busby 1994), FishGopher (see Wiley and Peterson 2004p. 92), and MaNIS (Stein and Wieczorek 2004)). Most data exchange initially was in support of taxonomic research, such as the description of new taxa, the writing of floras and faunas, and for writing monographs. Over time, demand has grown for biological data to be used for other purposes - for example for species distribution modelling (Longmore 1989, Peterson et al. 1998), biogeographic analysis and regionalisation (Thackway and Cresswell 1992), phylogenetic studies (Hamilton 2013), and conservation analysis (Ponder et al. 2001, Graham et al. 2004).

The development and expansion of the Internet has been a major driver increasing demand for data from a wider audience, reflected in the development of aggregation initiatives (Chapman and Busby 1994, Soberon et al. 1996), including some with specific purposes in mind, such as for species distribution modelling (Stockwell et al. 2006). In 2001, the Global Biodiversity Information Facility (GBIF) was established (Edwards 2004, Lane 2005) with the aim of aggregating data from the world's biological institutions, initially focusing on specimen data from museums and herbaria, then grid-based data from conservation initiatives (Yesson et al. 2007, Landuyt et al. 2012), and data from observation initiatives and citizen science projects (Levatich and Padilla 2016, Mackay 2017). More recently, GBIF has incorporated growing volumes of ecological, genetic, and other data. GBIF now serves as a platform for aggregating all forms of evidence for the occurrence of any species in time and space.

The availability of these data has provided opportunities for researchers and data practitioners to use biodiversity data in new ways (Chapman 2005c). Despite community efforts to standardise data transfer and delivery, this sharing of data and uses has identified previously unrecognised issues with the data and its quality for downstream applications (Rowe 2005, Beck et al. 2014, Maldonado et al. 2015). In 2013, TDWG established the Data Quality Interest Group (Saraiva and Chapman 2013) to examine and make recommendations on ways the data may be improved and documented so that users would have greater certainty with respect to the quality of the data available for their particular use. In 2015–2016, GBIF coordinated expert task groups on data fitness for use in three research disciplines: agrobiodiversity (Arnaud et al. 2017), species distribution modeling (Anderson et al. 2015), and alien and invasive species (McGeoch et al. 2016). Each of the expert groups was tasked to explore and to summarise the data demands and the data functionality improvements relevant for their respective research tasks.

The TDWG Data Quality Interest Group (DQIG) identified four main aspects that needed to be addressed in order to advance towards assessing and enhancing data quality and fostering its broader use:

a) **a data quality framework**, to provide a rigorous means to describe concepts of data quality, theoretical guidelines for data quality assessment and enhancement using a fitness for use approach, and a common means for describing data quality needs and analyses of the fitness for use of data for particular needs;

b) **data quality use cases**, to determine data quality profiles according to particular needs of the user, and lead to an understanding of the biodiversity concepts on which data quality assessments should focus;

c) **data quality tests and assertions**, to describe in a consistent manner what is being tested and under which assumptions/parameters, and so define the expected result that any application of the tests should be able to provide. Tests need to be uniform across the community, so that, regardless of the implementation, a given test should always return the same result given the same inputs and parameters.

d) **implementations of the tests**, to illustrate the ways in which the tests are applied and the specifics of how the results of the tests are presented. The internal details of how a specific test is performed are largely left up to the implementor. The test definitions and specifications are also deliberately agnostic concerning implementation languages and the software framework within which the tests are executed (e.g., frameworks that group data by unique values for testing, or data pipelines that operate record by record on multiple separate parallel test pipelines). Implementations need not necessarily be uniform across the community, but common specifications need to be followed, and combinations of tests must follow consistent data quality profiles with a standard form of reporting the results of tests in order to produce a common community of practice (as discussed below). Implementations of tests must support the expectations of consumers of data quality reports, so that the same test run on the same data by different tools will return the same result.

Following the four aspects described above, the DQIG initially established three Task Groups (TG), one to develop a framework for the assessment and management of data quality using a fitness for use approach (Veiga et al. 2017); a second to gather and classify user stories to form contextually themed use cases, such as species distribution modeling, agrobiodiversity, and invasive species; and a third to define a core set of standardised tests and associated assertions based on Darwin Core terms (Wieczorek et al. 2012). In this paper, we present an example implementation of the tests as a proof of concept of their applicability, using a subset of data quality tests implemented as Kurator (Morris et al. 2018) workflows. The work of these three Task Groups and of the TDWG Darwin Core Maintenance Group (Wieczorek 2006) later identified the need for a fourth Task Group to be established to develop controlled vocabularies of values for Darwin Core terms.

3. A Fitness for Use Framework

It is essential to have a formal way of talking about, describing and documenting data quality — the needs of data consumers, the tests that can evaluate data against those needs, and the results of such tests on sets of data. Formality is particularly important to enable different implementations of data quality tests to make consistent assertions in consistent and comparable ways. The first of the three Task Groups (largely the work of Veiga 2016), produced a framework for formal description of data quality (Veiga et al. 2017).

For any research question, or in order to test any specific hypothesis, it is important to identify and access the subset of globally and digitally available biodiversity data that are suitable for the purpose and meets the required thresholds of quality, precision, and accuracy. The process of discovering which fraction of data is suitable for addressing a given question is what we term 'data quality assessment'. Quality improvement (through data proofing and standardization) can be carried out to increase the amount of data that are fit for use. This process is what we are calling 'data quality management'. Judging which fraction of the data is fit for use (assessment) and how to improve the fraction that is not immediately fit for use (management) is not trivial and can only be carried out in the context of a clear definition of data quality needs.

Data quality assessment and management are highly dependent on context. For example, a piece of data may be fit for use in a specific context because it has consistent coordinates and accurate scientific name, but the same piece of data may not be fit for use in another context because the event date is not complete. Making this piece of data fit for the second use requires an action, i.e., completing the event date value.

The Fitness for Use Conceptual Framework (FFU Framework) provides a formal structure and a process for defining the criteria for determining that data are fit for use in a given context. This formalization separates the concerns of clear descriptions of data quality needs, descriptions of tools that act on data against those needs, and the content of data quality reports. The formalization specifies both the elements required to express a particular data quality need, and explicit connections between assertions made in data quality reports and data quality needs. To do this, the Framework provides a formal language and structures for describing data quality needs (through analysis of use cases), tools for assessing data quality, and data quality reports. The framework uses the concept of a data quality 'Profile' which defines the components necessary for assessment and management of the data in the context of a use case (Veiga et al. 2017). Based on the Profile, a set of data quality reports can be generated to describe the current status of the quality of a specific piece of data (a single record or set of records). Each resulting 'Report' comprises a set of data quality assertions generated by one or more data quality 'Solutions' (i.e., methods and mechanisms used to meet the data quality needs).

The following subsections use the conceptual framework to describe the processes of: (1) defining a data quality Profile to document data quality needs in a given context; (2) describing data quality Solutions to meet those data quality needs; and (3) structuring data

quality Reports for presenting the current status of quality of a data resource, according to a Profile, and thus the status of that data resource with regard to some data quality need.

3.1. Data Quality Profiling

Due to the idiosyncratic nature of the concept of 'quality', it is essential to understand what 'data fitness for use' means from the data user's perspective. We can not assert that some data resource has quality in the abstract, we can only assert that it has quality in regard to some data user's need for that resource, if it is fit for a particular use. It is understanding the user's needs in a systematic way, and expressing formally what attributes the data must have to be fit for a user's purpose, which enables data quality assessment and management (Chrisman 1991, Strong et al. 1997).

A data quality profile can be defined by following five steps (Fig. 1): (1) defining a use case; (2) defining the valuable Information Element (IE); (3) defining a data quality measurement policy (consisting of a set or Meaures with additional context); (4) defining a data quality validation policy (consisting of a set of Validations with additional context); and (5) defining a data quality improvement policy (consisting of a set of Improvements with additional context), which when we get to the report layer will produce proposals for Amendments*2 (Veiga et al. 2017).

Figure 1.

Fitness, Data, and Use components used by the FFU Framework, where the data quality (DQ) profile describes the fitness for use (or purpose) of some data. The use is expressed as a use case, the data are an identified set of information elements with value for that use, and the fitness can be expressed through descriptions of how to measure the fitness, how to validate whether the values in information elements meet the needs of the use, and what means could improve data that are not fit for the use, but might be able to become so.

A research question concerning change in the distribution of species over time (which we could phrase as a use case) will involve some aspect of quality of data with regards to

time. The researcher may say, a given set of occurrence records have quality for my use if the dates of observation are known to a resolution of one year or less. From this statement of a data quality need, we may identify dwc:eventDate*1 as a valuable information element, and can express "one year or less" as a set of specific measurement, validation, and improvement tests. We could assert that a measurement is needed to assess the duration of the interval expressed in the value found in a single record's dwc:eventDate, and that a positive validation test reflects the duration of the dwc:eventDate in a single record as less than one year (with a data quality dimension of "Resolution" - see definition in Suppl. material 1). We could also express ways of potentially improving the data by, for example, proposing an amendment to the data by populating an empty dwc:eventDate with the combination of values in dwc:day, dwc:month, and dwc:year. We could then include these tests in a data quality Profile for this research use and assert that for quality control for this use, all data in a set of multiple records must have all individual records passing the validation test of a dwc:eventDate with a duration of one year or less, and that all records failing this test are excluded. Or for quality assessment, we could express a measurement of what portion of the records in a set pass this validation test, and (combining with other tests in this Profile) assert that 40% of the records in some data set are fit for this use.

Explicit in the framework is the idea that data do not have quality, except in the context of some use. A data quality Profile tells us which tests (measurements, validations, amendments) are needed to assess data quality or perform data quality control for some use.

3.2. Data Quality Solutions

Data quality profiles define the requirements of a data user with respect to measuring, validating, and improving the quality of the data. Policies themselves do not improve data, so, for the implementation and application of profiles in an organization, concrete means must be defined for applying those policies. This is the domain of Data Quality Solutions. Data quality solutions carry the same tripartite division of measurement, validation, and improvement from the profile into the definition of specifications and mechanisms, as illustrated in Fig. 2. A Specification asserts how a test is to be implemented, a Mechanism is something (usually software) that provides that implementation. (Veiga et al. 2017).

Figure 2.

Components of a data quality (DQ) profile and of the data quality solutions in the FFU Framework.

For example, for any particular validation test, we can provide a specification of what that test needs to do (a specification detailed enough for a software developer to provide an implementation), and then separately assert that one or more independent software packages contain mechanisms capable of providing that test as part of a data quality Solution. Specifications should describe the test in pseudo-code in sufficient detail for unambigous implementation, and must make explicit those assumptions that may be glossed over in the policy. A policy may state, for example, that for a year in a single record to have quality for some use, then it must have a value between 1600 and the present year. A specification would go further in providing guidance to an implementer of the test in covering the expected behavior of error conditions (what to do if the year is blank, or if it doesn't contain a value interpretable as an integer representing a year), and what values the test is to return if the year is in range or if it is not.

3.3. Data Quality Reporting

Data quality reporting documents the status of the quality of a data resource according to a data quality profile as defined under section 3.1. A data quality report is composed of five components (from Veiga et al. 2017): (1) a context; (2) data resource; (3) data quality measures; (4) data quality validations; and (5) data quality amendments, as illustrated in Fig. 3. The tripartite theme of measurements, validations, and improvements/amendments is carried across the entire framework. This lets us tie a particular validation test result in a data quality report back to the mechanism that ran the test, back to the validation policy for the use case being tested. The formal concepts of the framework provide for both human readable and machine readable reports, where software can apply quality control for some use to a dataset relating a data quality report to a specific use case. The formal specification of information elements and validations, measures, and amendments, also allows for software implementations that understand how to apply the appropriate set of tests to the relevant data elements based on a selected data quality profile.

Figure 3.

Components of a data quality (DQ) profile, data quality solutions, and data quality reports used in the FFU Framework.

With these five components, we are able to build a data quality report that presents the status of the quality of a data resource in a given context of a use case and its respective profile, as illustrated in Fig. 4. The framework, by expressing a structure for the report, allows for the generation of multiple forms of human readable reports from machine readable report results. For example, in the Kurator project (Morris et al. 2018), color coded (human readable) spreadsheets of flat Darwin Core data translate the machine

readable specification of the information elements defining which cells to highlight, and the specification of validation test result values to know what color highlighting to use (e.g., red for validation failures).

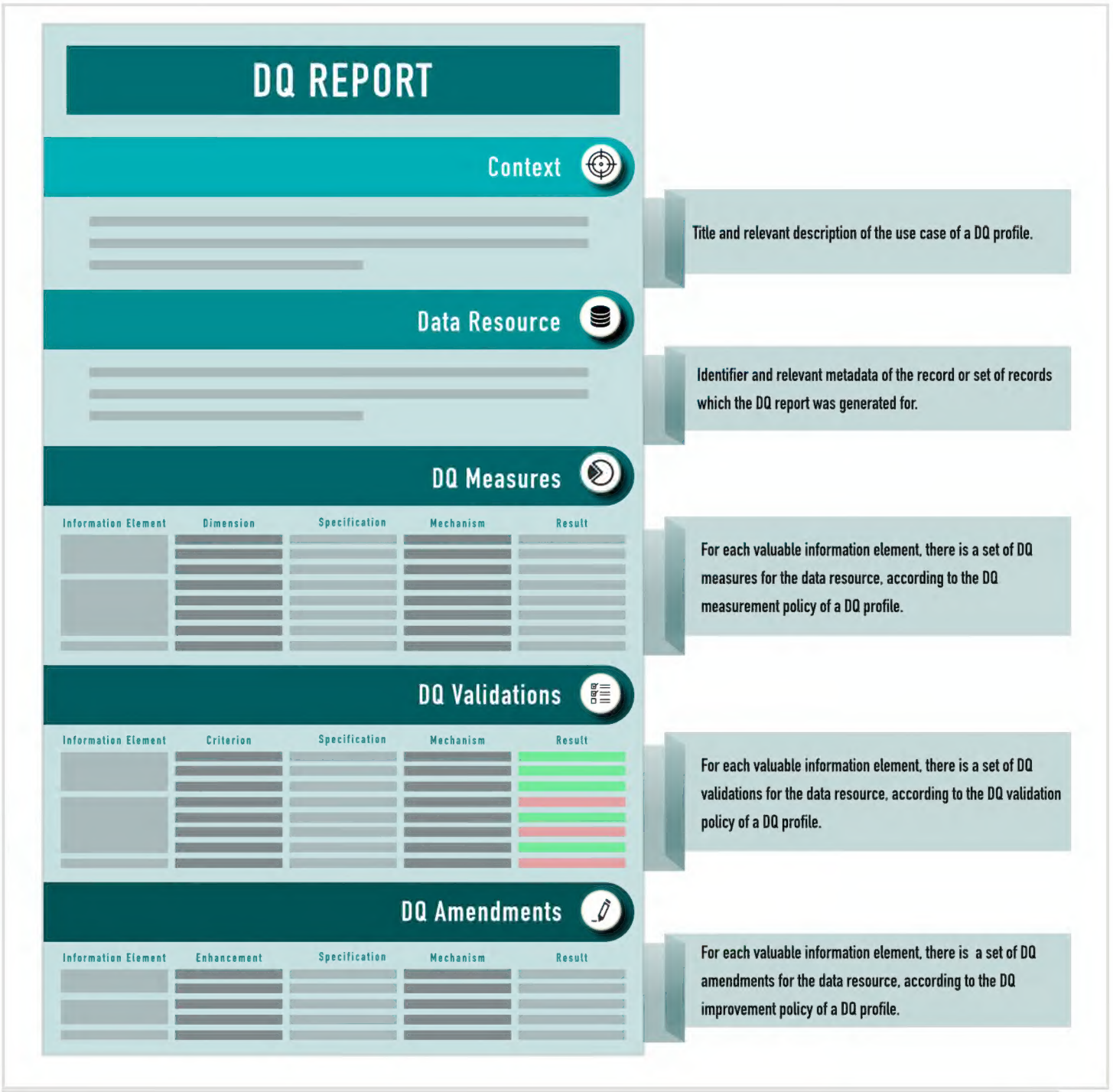
DQ REPORT
Context
Title and relevant description of the use case of a DQ profile.
Data Resource
Identifier and relevant metadata of the record or set of records which the DQ report was generated for.
DQ Measures
Information Element
Dimension
Specification
Mechanism
Result
For each valuable information element, there is a set of DQ measures for the data resource, according to the DQ measurement policy of a DQ profile.
DQ Validations
Information Element
Criterion
Specification
Mechanism
Result
For each valuable information element, there is a set of DQ validations for the data resource, according to the DQ validation policy of a DQ profile.
DQ Amendments
Information Element
Enhancement
Specification
Mechanism
Result
For each valuable information element, there is a set of DQ amendments for the data resource, according to the DQ improvement policy of a DQ profile.

Figure 4.

Example of a data quality report structure.

3.3.1 Context

The choice of data quality profile under which a data resource is examined may result in different assertions about the quality of that data resource. For one data quality profile, a data resource may have quality, for another profile, the same data resource may not. For example, the same record may have sufficient coordinate resolution for a given use case but insufficient coordinate resolution for another use case. A study of phenology may require temporal resolution of occurrences down to a day, while a study of long-term changes in species distribution may be satisfied with resolution down to a year.

3.3.2 Data resource

A data resource is an instance of assembled facts. This instance can be a single record (e.g., a Darwin Core formatted record, a row in a spreadsheet, a citizen science observation recorded in a mobile application) or a set of records (e.g., an institutional dataset, aggregated data from different sources, a result obtained from a specific query in a data portal). The quality of the data resource in the context of a use case is determined by three components:

1. **Data quality measures** - present assertions of measures of the data quality of a data resource in accordance with a given specification and mechanism (Fig. 5). The measures are structured as data resource, information element, dimension, specification, mechanism, and result.

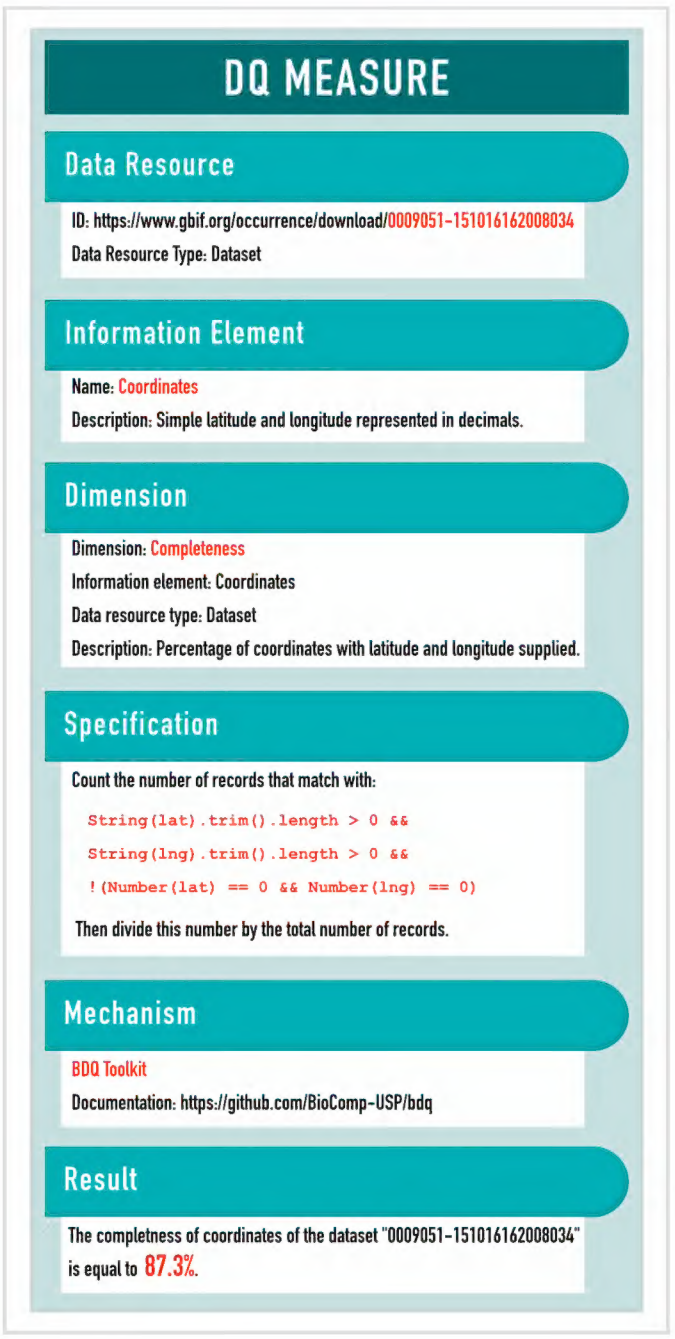
DQ MEASURE
Data Resource
ID: https://www.gbif.org/occurrence/download/0009051-151016162008034
Data Resource Type: Dataset
Information Element
Name: Coordinates
Description: Simple latitude and longitude represented in decimals.
Dimension
Dimension: Completeness
Information element: Coordinates
Data resource type: Dataset
Description: Percentage of coordinates with latitude and longitude supplied.
Specification
Count the number of records that match with:
String(lat).trim().length > 0 &&
String(lng).trim().length > 0 &&
!(Number(lat) == 0 && Number(lng) == 0)
Then divide this number by the total number of records.
Mechanism
BDQ Toolkit
Documentation: https://github.com/BioComp-USP/bdq
Result
The completness of coordinates of the dataset "0009051-151016162008034" is equal to 87.3%.

Figure 5.

Example of a data quality measure.

2. **Data quality validations** - present a set of assertions that report if a data resource is compliant with a specific criterion, according to a given specification and mechanism (Fig. 6). The validations are structured as data resource, information element, criterion, specification, mechanism, and result.

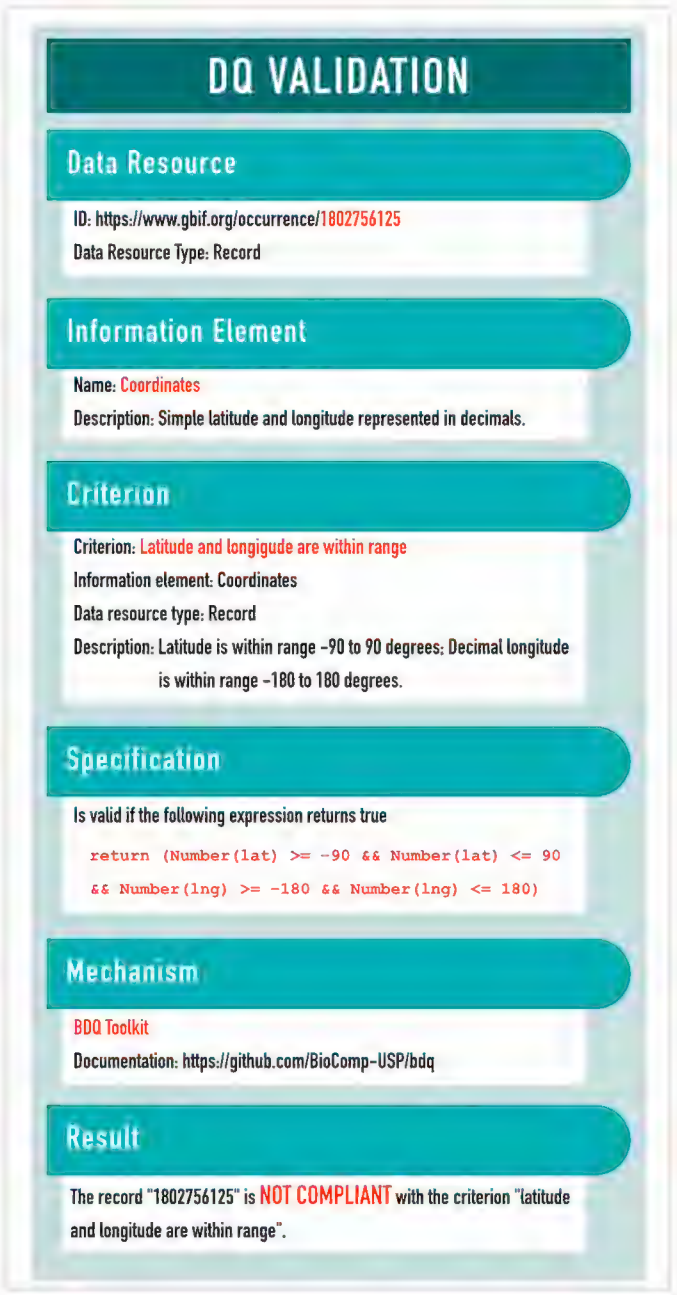
DQ VALIDATION
Data Resource
ID: https://www.gbif.org/occurrence/1802756125
Data Resource Type: Record
Information Element
Name: Coordinates
Description: Simple latitude and longitude represented in decimals.
Criterion
Criterion: Latitude and longigude are within range
Information element: Coordinates
Data resource type: Record
Description: Latitude is within range -90 to 90 degrees; Decimal longitude is within range -180 to 180 degrees.
Specification
Is valid if the following expression returns true
return (Number(lat) >= -90 && Number(lat) <= 90 && Number(lng) >= -180 && Number(lng) <= 180)
Mechanism
BDQ Toolkit
Documentation: https://github.com/BioComp-USP/bdq
Result
The record "1802756125" is NOT COMPLIANT with the criterion "latitude and longitude are within range".

Figure 6.

Example of a data quality validation.

3. **Data quality amendments** - present a set of suggested amendments for improving the quality of a data resource according to a given specification and mechanism (Fig. 7). The amendments are structured as data resource, information element, data enhancement, specification, mechanism, and result (the amendment in the result is linked to the concept "improvement" in the data quality profile, however we often informally refer to amendments across the board in this layer).

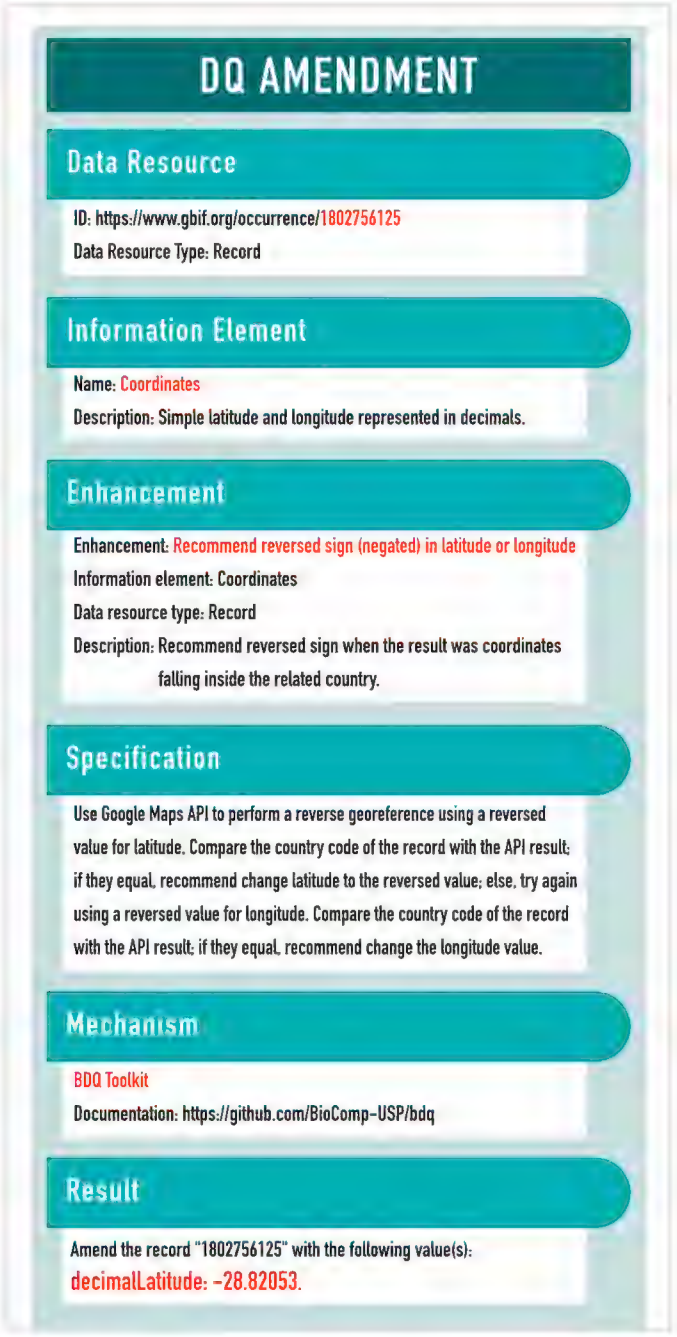
DQ AMENDMENT
Data Resource
ID: https://www.gbif.org/occurrence/1802756125
Data Resource Type: Record
Information Element
Name: Coordinates
Description: Simple latitude and longitude represented in decimals.
Enhancement
Enhancement: Recommend reversed sign (negated) in latitude or longitude
Information element: Coordinates
Data resource type: Record
Description: Recommend reversed sign when the result was coordinates falling inside the related country.
Specification
Use Google Maps API to perform a reverse georeference using a reversed value for latitude. Compare the country code of the record with the API result; if they equal, recommend change latitude to the reversed value; else, try again using a reversed value for longitude. Compare the country code of the record with the API result; if they equal, recommend change the longitude value.
Mechanism
BDQ Toolkit
Documentation: https://github.com/BioComp-USP/bdq
Result
Amend the record "1802756125" with the following value(s):
decimalLatitude: -28.82053.

Figure 7.

Example of a data quality amendment.

4. Collecting User Stories that lead to Data Quality Profiles

The Data Quality Use Cases Task Group of the TDWG Data Quality Interest Group was tasked to develop a set of use cases that are being applied by agencies and user communities to select records, and/or data sets, for particular purposes (https://www.tdwg.org/community/bdq/tg-3/). The use cases were placed into a use case library, which allowed us to develop data quality Profiles, and from these identify a core set of tests for these Profiles.

The strategy proposed was:

- Document use cases in a structured format based on the FFU Conceptual Framework (Veiga et al. 2017).

- Place a use case template in a collaborative editing environment for completion and discussion.
- Contact government and conservation agencies and user communities to establish and document use cases where they assess data fitness for use.

4.1 Use Case Selection

Use cases were collected by a number of methods to maximise responses. Based on the framework described above (Veiga et al. 2017) a spreadsheet was developed, as well as a simpler Google Form. Lead authors of papers published using data accessed via the Atlas of Living Australia (ALA) were contacted and asked to contribute their research data use cases, and six papers describing fitness for use determination were sent to the ALA Data Quality group. Fitness for use and quality checking information from these papers were extracted and transferred across to the use case library.

Three face-to-face interviews with researchers were also conducted. Interviewees were asked to describe their research projects and the checks they apply to the downloaded data before use. Questions asked approximated those from the Google Form, with additional information requested where details were lacking.

This study included data from 26 submitted use cases, as well as two example use cases based on general criteria for studies of ecological niche modelling and ecological gap analysis.

4.2 Mapping the Data Collected to the Use Case Library

To extract information from the submissions, the use cases were broken down for each mention of data quality criteria utilized. Using the pre-built skeleton for the use case library, these mentions were mapped onto relevant criteria, dimensions and information elements in accordance with the FFU Conceptual Framework.

Information regarding the source of the use case and the application for which the data was being used, was collected to allow for additional analysis. Use cases were sorted into application groups based on the published papers and on how the use case had been described. Categories included distribution modelling, database entry, and species list development.

4.3 Use Case Library Analysis

From other assessments of data quality and fitness for use studies (e.g., Arnaud et al. 2017), analyses were based on the frequency with which certain information elements, dimensions, and criteria of data quality were used. A similar methodology was employed in this study. The number of use cases associated with each information element, dimension and criterion was counted, as was the number of information elements, dimensions and criteria associated with each use case, and the mean, median and mode determined for each. Note that definitions for individual criteria, information elements, and dimensions can

be found in Suppl. material 1. Basically, criteria describe the acceptable levels of data quality, information elements are terms that represent the content, and dimensions are the measures of the quality of the criteria (Veiga et al. 2017). Each criteria and dimension were also associated with a fundamental criterion and fundamental dimension respectively, and the number of use cases associated with each was also assessed (Suppl. material 2).

4.4 Quality and Area of Interest

Each of the 257 criteria indicated was classified as either a 'quality' criterion or an 'area of interest' criterion. Criteria such as “precision should be between 0 and 1” and “coordinates must not be generalised” were classified as 'quality' based, whereas criteria such as “record must not be classified as a machine observation” or “coordinates must fall within forest or woodland classed land cover areas” were classified as 'area of interest criteria'. 'Area of interest' classifications were more subjective, being based on information provided about how the data were used, and not directly indicated by the data user. Of the 257 criteria, 211 were classified as 'quality' based, and 46 as 'area of interest'. The majority of criteria used to assess fitness for use assessed an aspect of the record that was not necessarily specific to the project, and as such could be improved for the majority of applications and uses.

4.5 Criteria

Criteria are statements that describe acceptable levels of data quality. The most commonly used criteria included: coordinates are complete, the scientific name must match a given reference list, the coordinates must be within a given range, the basis of record is well formed, and all the coordinates in a dataset are complete (Fig. 8).

Most use cases utilized 10 criteria when filtering biodiversity data. The mean number of criteria used in each use case was 13.11, the mode 11 and the median 8. When the extreme use case based on 101 criteria was excluded from the calculations, the mean was 9.85 criteria. Fig. 9 shows the number of criteria (orange lines) associated with each use case.

To obtain a more useful representation of the usage of criteria, each criterion was given a fundamental criterion. A fundamental criterion is a grouping of criteria based on a similar metric. For example, the fundamental criterion "DQ measure must be in the range" includes all criteria based on a data range, such as "Precision should be between 0 and 1" and "Coordinates must be located within the expected region". The list of fundamental criterion were taken from Table F of the supplementary material ("*S1 Supporting Information – Case Study*") in Veiga et al. 2017. This grouped several criteria together and showed that the fundamental criterion "Data Quality measure must be in the range" is the predominant criteria used to assess fitness for use (Fig. 10).

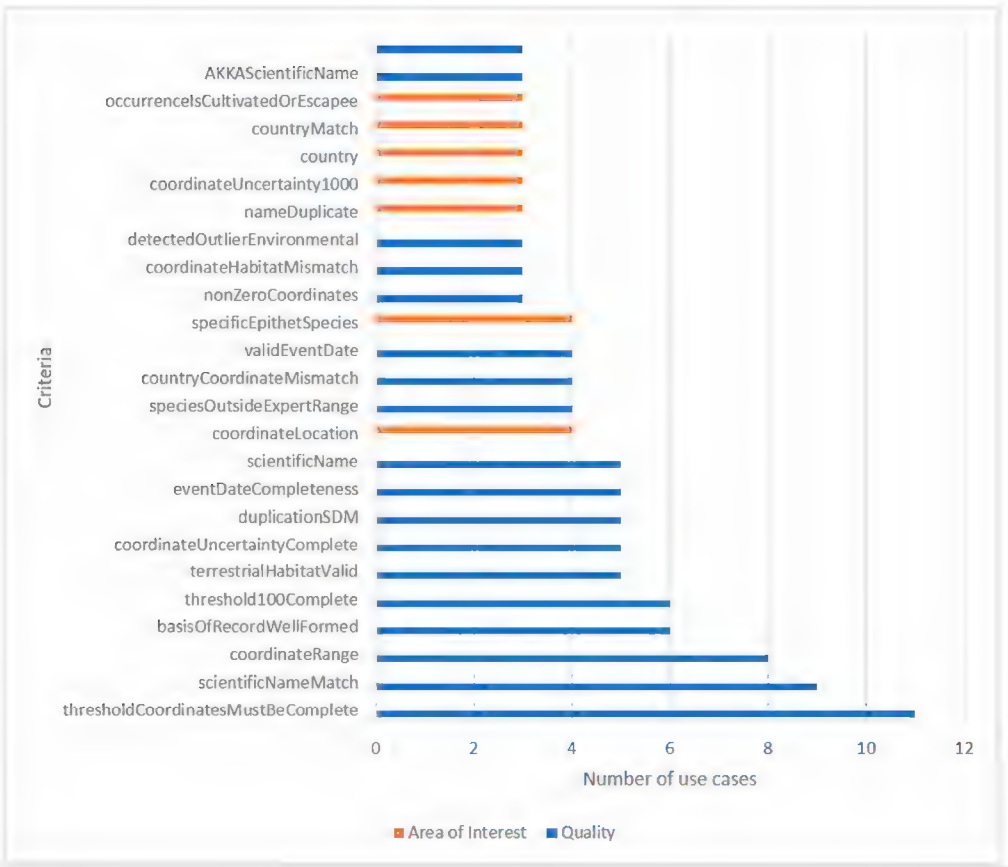

Figure 8.

Number of use cases associated with each criterion (see a full list of criteria in Suppl. material 2). There were 232 criteria that have at least one, but fewer than three use cases and are not shown. Criteria that are largely based on an Area of Interest are shown in Orange, and those on Quality Criteria in Blue.

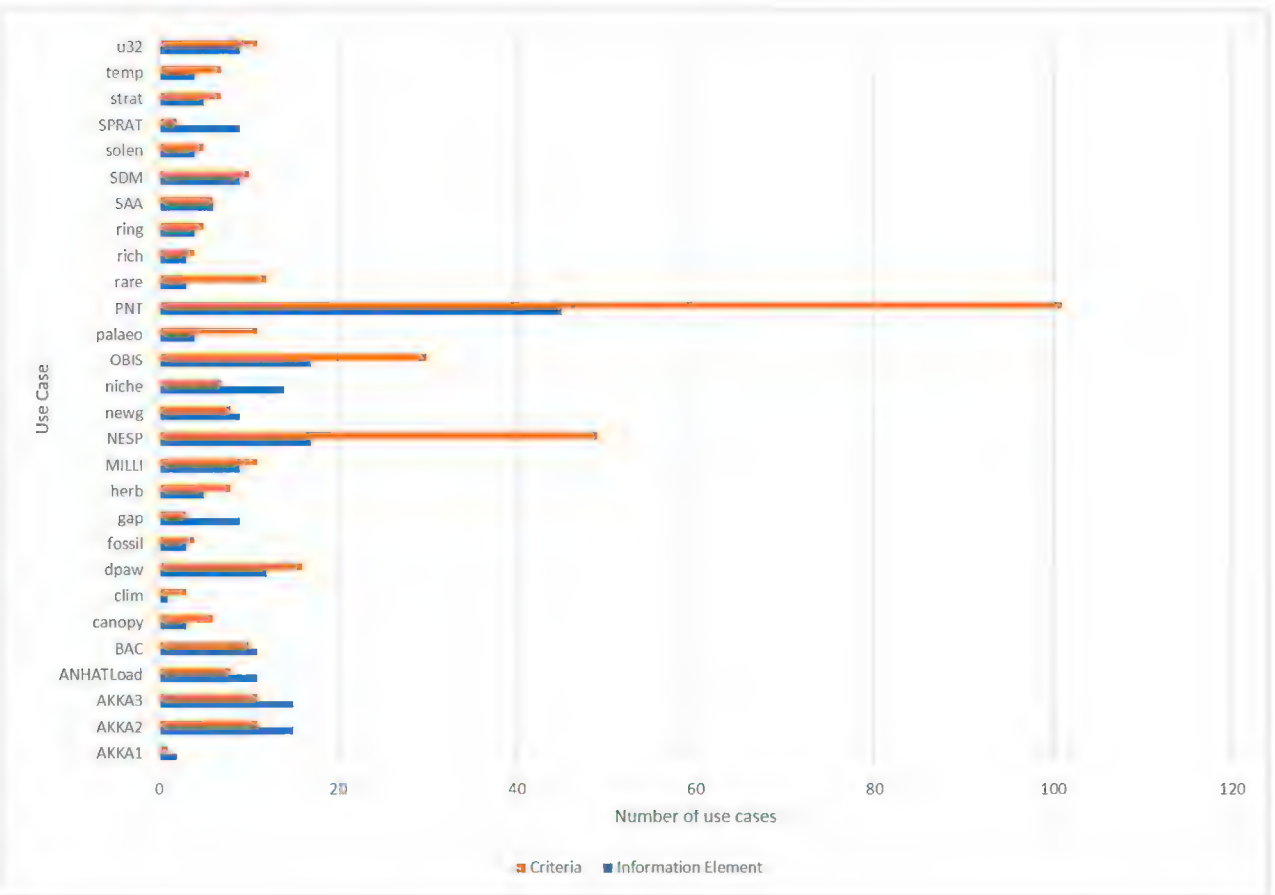

Figure 9.

Number of criteria (orange) and information elements (blue), used by each use case for data download. A list of criteria and uses cases (abbreviated here) can be seen in Suppl. material 2.

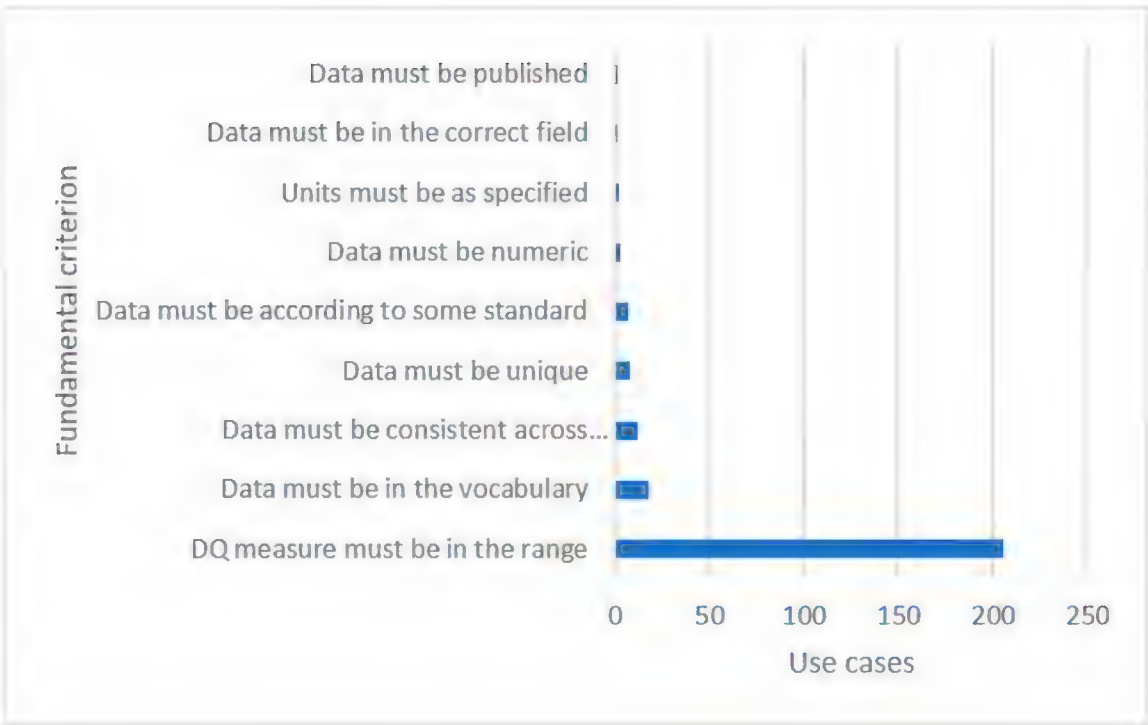

Figure 10.

Number of use cases associated with each fundamental criterion (see Suppl. material 2). As each use case can be associated with multiple criteria, use cases may be double counted when criteria are combined to form fundamental criterion.

4.6 Information Elements

The most used information elements (i.e. terms that represent relevant content) indicate areas where the most effective data quality improvements could be made. The analysis of the distribution of information elements indicated that there were a number of information elements that had a higher number of use cases associated with them than other information elements. These included the Darwin Core terms dwc:coordinates, dwc:decimalLatitude and dwc:decimalLongitude (of which coordinates are composed), dwc:scientificName, dwc:eventDate, and dwc: country (Fig. 11). The remaining information elements were all used by fewer than one third of the documented use cases.

The mean, mode and median number of information elements per use case was approximately nine (9.214, 9, 9 respectively), with the greatest number of elements associated with a single use case being 45. This implies that determining fitness for use of a record or dataset is most commonly based on nine information elements. Fig. 9 shows the number of information elements (blue lines) associated with each use case.

The information element analysis clearly showed the importance of locality information, indicating that improving this aspect of a record could have a significant impact on the overall quality of the record. Coordinates, decimalLatitude and decimalLongitude all contain location information, and their usage by almost every use case included in the analysis shows that locality information is consistently one of the most crucial components in determining fitness for use. When broken down into categories according to the Darwin Core quick reference guide (Biodiversity Information Standards (TDWG) 2018), the prevalence of location information was reinforced. The number of use cases associated

with a Darwin Core term in the Location category was almost equal to the number of use cases associated with a term in all the other categories combined (Fig. 12).

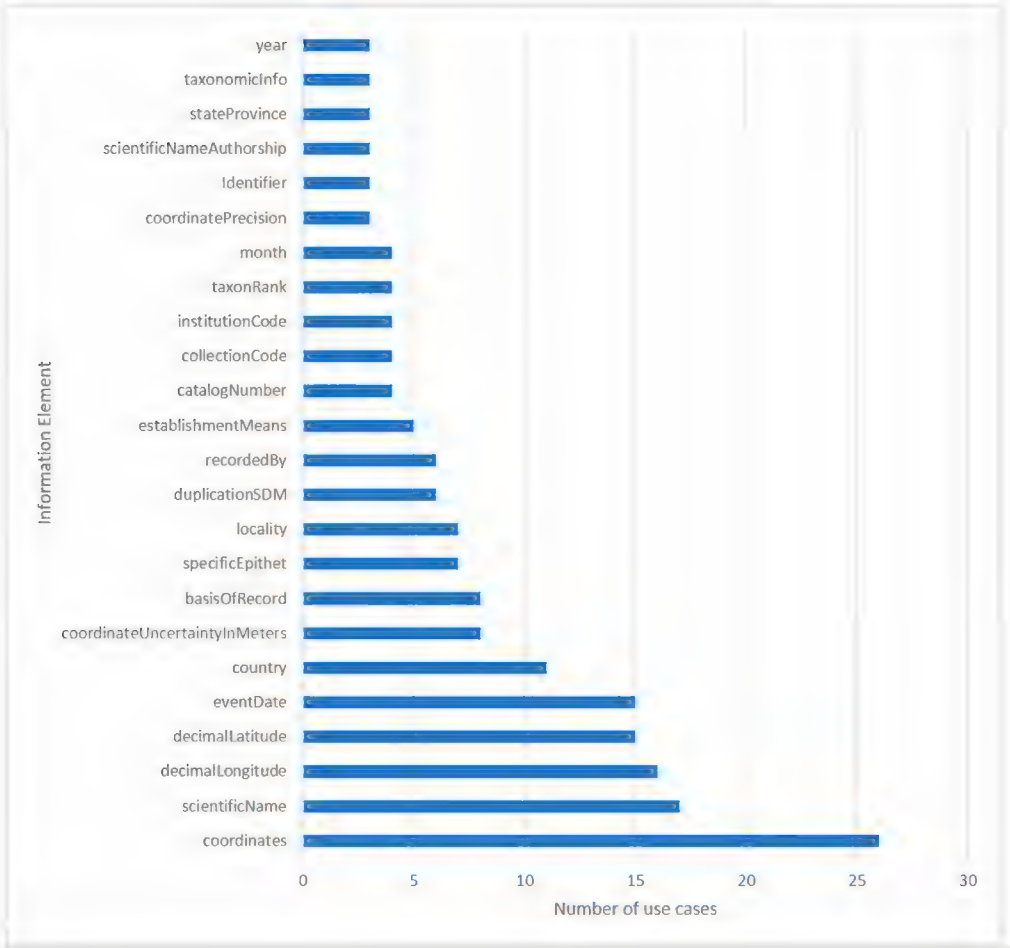

Figure 11.

Number of use cases associated with each Information Element (ie) (see Suppl. material 2 for list of Information Elements). Information Elements that had fewer than three associated use cases are not shown.

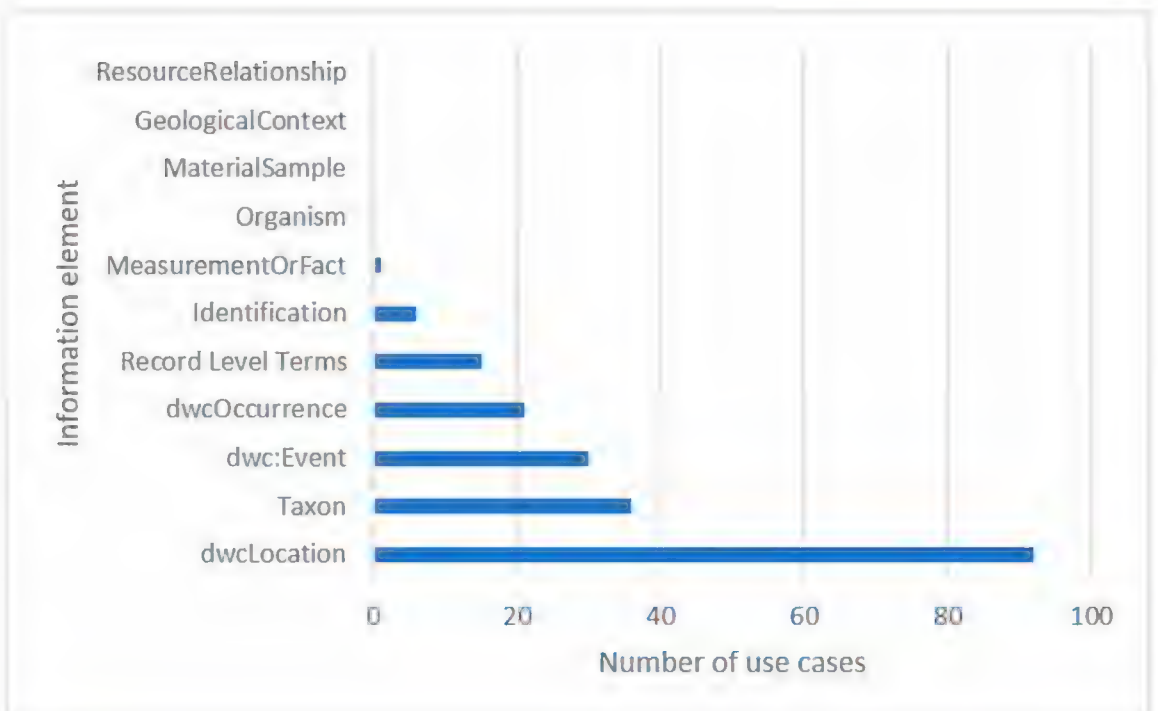

Figure 12.

Number of use cases associated with Information Elements (ie) in each Darwin Core quick reference guide (Biodiversity Information Standards (TDWG) 2018) category.

4.7 Dimensions

A dimension is the measureable quality of a criterion. To be able to effectively improve the quality of the locality information available, an understanding of the importance of aspects of locality is required. As with the information elements, there were a number of key dimensions that were used more frequently than others (Fig. 13). A measure of the value of the coordinates (d:coordinatesValue) was the most frequently used dimension. Not all use cases were concerned with the value of the coordinates, however, indicating that whilst this is important, it is not the only dimension used to determine fitness for use.

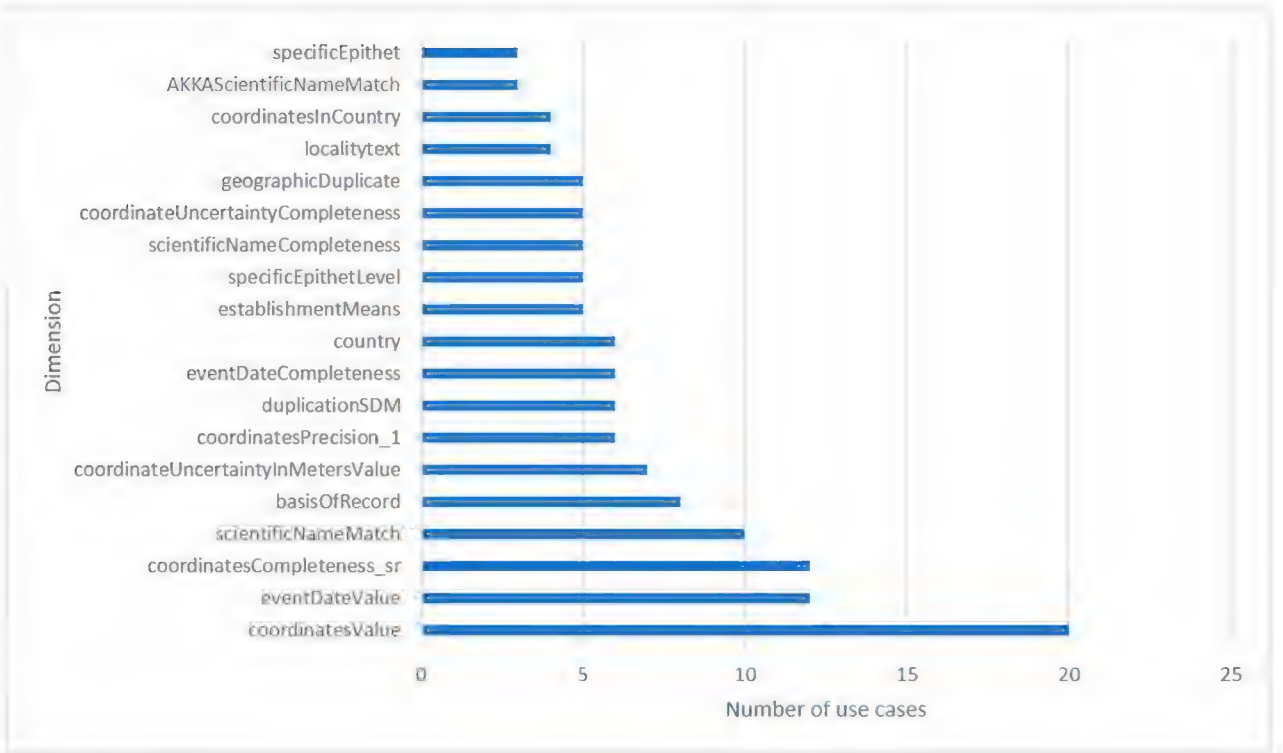

Figure 13.

Number of use cases associated with each Dimension (for a full list of dimensions, see Suppl. material 2). Dimensions that had fewer than three associated use cases are not shown.

Each use case was associated with one of nine fundamental dimensions. 'Value' and 'Completeness' were the most commonly used fundamental dimensions (75%), indicating that data users are often requiring certain pieces of information as a baseline, and assessing some aspect of the value, be that it lies within a given range, or must be of a given set of values. Fig. 14 shows the distribution of the number of use cases associated with each fundamental dimension.

5. Data Quality Tests and Assertions: a Data Quality Solutions Library

As noted above, evaluation of the 'quality' of data is dependent on the use to which data will be applied. There are however generic and programmatically testable issues within data records that will assist in evaluating fitness for use. For example, a record may contain a latitude that doesn't match the supplied country. Experience, and the analysis of use cases above, suggests that a standard suite of tests and resulting assertions would be

useful in evaluating the 'fitness for use' of occurrence records for many different uses. A basic suite of tests based on terms of the international standard, Darwin Core (Wieczorek 2006, Wieczorek et al. 2012) seemed to be a practical first step in assessing the fitness for use of biodiversity data records. Darwin Core was an obvious scope for the tests as this standard is by far the most used for the documentation and exchange of occurrence records. Such a suite of tests/assertions could be used in applications from those who collect the data through to those who want to use the data. The TDWG Data Quality Interest Group identified the need for a Task Group to take responsibility for compiling this core suite of tests that identified record issues and attempted to improve and report on these issues.

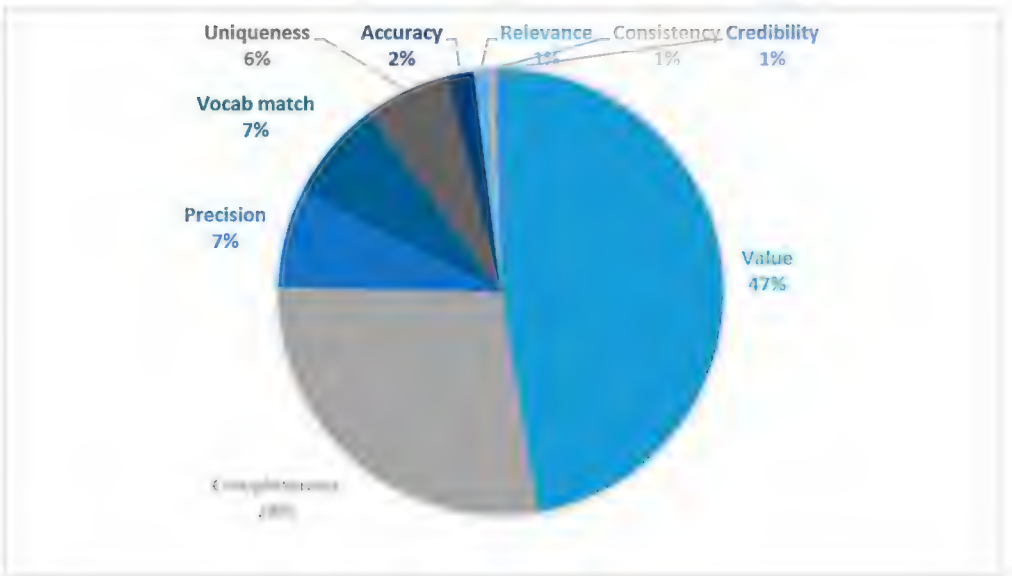

Figure 14.

Distribution of the number of use cases associated with each fundamental dimension (see Suppl. material 1 for definitions of these dimensons).

5.1 Context for the Tests and Assertions

The first step in this compilation was to review all the tests for addressing data quality, which were currently used by data publishers such as the Atlas of Living Australia (http://www.ala.org.au), Biodiversity Information Serving Our Nation (http://bison.usgs.gov), the Centro de Referência em Informação Ambiental (http://www.cria.org.br), the Global Biodiversity Information Facility (http://www.gbif.org), iDigBio (http://www.idigbio.org), and the Ocean Biogeographic Information System (http://www.iobis.org). This review resulted in over 200 tests with inevitable overlaps. The tests as currently implemented by these aggregators/publishers are 'negative' in the sense that they identify issues with the record/s. The tests can however be structured within the above FFU Framework as 'pass filters' (Quality Assurance in the sense of the framework Veiga et al. 2017 p.3) that filter sets of records down to just those that comply with all the requirements for quality under a selected Profile (e.g., a set of records that are compliant with the set of validations specified in some data quality Profile are fit for the use identified in that Profile). 'Negative' tests can be complemented with, or phrased as, their 'positive' counterparts that better fit the Data Quality Framework. Under the framework, a Validation reports its result values as COMPLIANT (which we can think of as a pass condition), and NOT_COMPLIANT (which

would be a fail or issue condition), but can also have a status indicating that prerequisites for the test were not met, something that may also be thought of as a fail or issue condition. Universal current usage and our experiences suggest that consumers of data quality reports focus on warnings or error messages (Quality Control in the sense of the Framework). Data quality control reports should thus emphasise the cases where Validations report that data are NOT_COMPLIANT (and that prerequisites are not met), rather than the hopefully much larger number of cases where Validations report that the data are COMPLIANT. That is, we would hope that the number of 'issue flags' would be far fewer than the number of 'pass flags' in any quality control report. Conversely, under Quality Assurance, record sets are filtered to exclude all records that have a value of NOT_COMPLIANT for any validation (which is part of the use case for the desired use of the data). For Quality Assurance, the data consumer's interest is likely the set of Validations for the use of the data at hand, rather than the individual issues.

Darwin Core terms can either be 'verbatim' where the values are effectively unbounded, or bounded by some constraints. Verbatim fields such as dwc:identifiedBy cannot easily be checked as new names must always be assumed possible. Terms such as dwc:decimalLatitude and dwc:decimalLongitude can however be easily checked for range bounds. This observation emphasized a known issue with many of the Darwin Core terms, i.e., that more controlled vocabularies (Section 7) were required for some terms to maximize data re-use. For example, for the term dwc:basisOfRecord it is recommended that the use of the terms be limited to "PreservedSpecimen", "FossilSpecimen", "LivingSpecimen", "HumanObservation", and "MachineObservation" but a review of the values in GBIF and the ALA for this term reveals over 2,000 distinct values. Such unconstrained values of many of the Darwin Core terms makes validation difficult to impossible. The review and formulation of a standard suite of tests within the Task Group has therefore lead to the formation of a Task Group on vocabularies of values under the Data Quality Interest Group (see https://www.tdwg.org/community/bdq/tg-4/).

A reasonable set of criteria for including a test in the standard suite of core tests includes: (1) tests should be simple and informative, imparting useful information about the status of one or more term values in the record; (2) tests should be relatively easy to implement to encourage broad application, from data collection to use evaluation; (3) tests should also have 'power' in the sense that we would generally avoid tests where the majority of records either passed or failed; and (4) the tests should examine terms identified as widely important for multiple use cases as in the analysis above. We also considered tests where we expect a "Fail" (NOT_COMPLIANT) on most records but where we intend to make a point (e.g., where the dcterms:type value is EMPTY) about the importance of a seldom populated term.

Ideally, assertions resulting from the tests should remain with the occurrence records. This is true for both Quality Assurance, where the test results provide provenance for the filtering and modification of the data set, and for Quality Control where the test results provide guidance on improving the quality of the data set. There are, however, workflows such as testing the mapping of data in some arbitrary schema onto Darwin Core for exchange, which may result in multiple cycles of testing and improvement where the test

results are of transient interest only. The test definitions themselves are largely agnostic about the workflow context within which they are used. Generally required criteria for amendments to correct or improve a record require a prior validation test of "Fail", a corresponding assertion that a change was made, and a subsequent re-validation that the amended value passed. For example, if the wrong sign on dwc:decimalLatitude was detected because of a mismatch with the value of the term dwc:country, the sign of dwc:decimalLatitude could be corrected and an assertion made that a proposal has been made to amend the value of dwc:decimalLatitude and, if this proposal is accepted, the revalidation test on dwc:decimalLatitude should report "Passed" (was COMPLIANT).

Missing (null) Darwin Core terms should not normally report NOT_COMPLIANT unless all taxon, all spatial, or all temporal terms are missing. The reason for this is that the number of Darwin Core terms that are completed can vary from three to over a hundred with different Darwin Core classes being supported by different biological communities and domains. Having a considerable number of what could be, in effect, false positives from missing Darwin Core terms or present but EMPTY, would devalue the tests and assertions. Thus, we excluded many possible validations of the presence of data in individual Darwin Core terms. If however, there is no information about any of the terms in three basic Darwin Core classes (Taxon, Event, Location), we consider fundamental data are missing and this is flagged by an assertion (NOT_COMPLIANT).

5.2 Test Descriptions and Parameters

After a comprehensive review of tests in use, data consumer needs from the use case analysis above, and the identification of gaps, 101 core tests were developed (developed as issues at https://github.com/tdwg/bdq/labels/Test, and exported to https://github.com/tdwg/bdq/blob/master/tg2/core/TG2_tests.csv). The test descriptions are included here as Suppl. material 4. It is the intent of the task group to bring these test definitions forward for ratification as a TDWG standard. For each test, the following standard information was compiled, covering both elements required by the framework, and additional metadata (Table 1).

Table 1.

Description of the terms used in the tests. A vocabulary can be found as Suppl. material 1. Fields required by the framework are noted with an (F), fields that extend the framework are noted with an (Ex), other fields are largely informative metadata.

Field	Description	Example
GUID	A globally unique identifier for each test, which allows software to uniquely identify each test (and in combination with parameter values, allows for specification of the expectations for the behavior of a test implementation).	e39098df-ef46-464c-9aef-bcdeee2a88cb

Field	Description	Example
Label	A standardised, human readable name of the test-assertion based on the template OUTPUTTYPE_TERMS_RESPONSE. These names were considered helpful for human-human communication and to assist with code implementation, maintenance and searches.	"VALIDATION_BASISOFRECORD_NOTSTANDARD"
Type (F)	Tests have been classified into one of three FFU Framework classes: VALIDATION (flags suspicious or invalid values in one or more Darwin Core terms); AMENDMENT (in terms of the framework, an IMPROVEMENT that will result in an AMENDMENT in a report, i.e., a change or addition to at least one Darwin Core term); and MEASURE (returns a numeric value, for the tests described here; all values are in the form of the number of tests that conform to a criterion). In addition, some tests are typed as NOTIFICATION (flags a potential issue where data remains valid, a concept outside the FFU Framework)	VALIDATION
Information Element Class	The Darwin Core class that the test relates to.	dwc:Taxon
Information Element (F)	The specific Darwin Core terms that the test takes as input.	For "VALIDATION_TAXON_AMBIGUOUS", dwc:taxonRank
Specification (F)	A concise description of the specification of the test for implementors, asserting the expected response including failure conditions in the form (for a VALIDATION) of: EXTERNAL_PREREQUISITES_NOT_MET if <condition>; INTERNAL_PREREQUESITES_NOT_MET if <condition>; COMPLIANT if <condition>; otherwise NOT_COMPLIANT.	For "VALIDATION_MONTH_NOTSTANDARD", "INTERNAL_PREREQUISITES_NOT_MET if the field dwc:month is EMPTY; COMPLIANT if the value of the field dwc:month is an integer between 1 and 12 inclusive; otherwise NOT_COMPLIANT"
Information Element Category	The information element dimension that the test refers to among Name, Space, Time or Other	For "VALIDATION_TAXONRANK_NOTSTANDARD", the Dimension is "Name"

Field	Description	Example
Data Quality Dimension (F)	A test will focus on one of the following scenarios based on the Data Quality Framework: "Completeness" (the extent to which data elements are present and sufficient); "Conformance" (Conforms to a format, syntax, type, range, standard or to the own nature of the information element); "Consistency" (agreement among related information elements in the data); "Likeliness" (low probability that values are real); "Resolution" (is sufficient detail present in the value/s - a measure the granularity of the data).	Completeness: "VALIDATION_TAXONID_EMPTY", Conformance: "VALIDATION_YEAR_NOTSTANDARD", Consistency: "VALIDATION_EVENTDATE_INCONSISTENT", Likeliness: "VALIDATION_COORDINATES_ZERO", Resolution: "VALIDATION_DATAGENERALISATIONS_NOTEMPTY"
Resource Type (F)	Whether this test examines a single record "SingleRecord" or a set of records "MultiRecord". Each VALIDATION acting on resource type SingleRecord is expected to be accompanied by a MEASURE counting compliance of that VALIDATION across a MultiRecord resource.	SingleRecord.
Warning type (Ex)	The nature of the flag associated with the test. Possible values are "Ambiguous", "Amended", "Incomplete", "Inconsistent", "Invalid", "Notification", "Report" and "Unlikely".	For "VALIDATION_FAMILY_NOTFOUND", the warning is "Invalid"
Parameter(s) (Ex)	Parameters that modify the behavior of the test, along with default values or links to source authorities	For "GEODETICDATUM_ASSUMEDDEFAULT": "bdq:sourceAuthority = (default = http://www.epsg.org/)". For "MINDEPTH-MAXDEPTH_OUTOFRANGE": "Default values bdq:minimumValidDepth = 0 and bdq:maximumValidDepth =11000"
Example	A concise example of the application of the test.	dwc:taxonRecord="sp." becomes dwc:taxonRank="species"
Source	The origin of the concept of the test.	Data Quality Interest Group Meeting during the TDWG 2018 Annual Conference in Dunedin, NZ.
References	One or more publications that relate directly to the test.	http://rs.gbif.org/vocabulary/gbif/rank.xml
Example Implementations (Mechanisms)	A link to one or more agencies that have an implementation of the test.	https://github.com/FilteredPush/event_date_qc

Field	Description	Example
Link to Specification Source Code	A link to reference code set that demonstrates the test.	https://github.com/FilteredPush/ event_date_qc/blob/5f2e7b30f8a8076977b2a609e0318068db80599a/src/main/java/org/filteredpush/qc/date/DwCEventDQ.java#L169
Notes	Additional comments that the Task Group believed necessary for an accurate understanding of the test or issues that implementers needed to be aware of.	For "VALIDATION_COUNTRYCODE_NOTSTANDARD", Locations outside of a jurisdiction covered by a country code should not have a value in the field dwc:countryCode.

Responses from each of the tests are expected to be structured data, not simple pass fail flags, including an assertion (which can form part of a data quality report or be wrapped in an annotation) with three components:

1. **Value** is the returned result for the test, i.e. numeric for measures, a controlled vocabulary (consisting of exactly COMPLIANT or NOT_COMPLIANT) for validations, and a data structure (e.g., a list of key value pairs) for proposed changes in amendments.
2. **Status** provides a controlled vocabulary, metadata concerning the success, failure, or problems with the test. The Status also serves as a link to information about warning type values and where in the future, probablistic assertions about the likeliness of the value could be made.
3. **Remark** supplies human-readable text describing reasons for the test result output.

We are aware of the centrality of the work of the TDWG Annotations Interest Group (https://github.com/tdwg/annotations) as to how the test results are reported against records. Test results structured with these three components can be readily wrapped in the body annotation document that follows the W3C Web Annotation Data Model (Sanderson et al. 2017), along with metadata from the Framework to describe which test is being reported upon, and metadata within the target of the annotation to describe which data resource is being annotated, and the state it was in at the time of annotation.

The set of core tests form a baseline Data Quality Profile thought to be broadly applicable based on our analysis of the frequency of populated terms in the wild and the use cases examined above, but we anticipate use case and domain-specific tests will be required. For example, for the marine environment, a test such as “minimum depth in meters is greater than indicated on GEBCO chart” may be appropriate. Similarly, for paleontological records a test "If dwc:basisOfRecord="FossilSpecimen" then at least one of the terms group, formation, member, or bed in a single record must contain a value" would be logical. We would urge those domains/communities needing additional tests to use the template and vocabulary defined here to ensure that a standard description of the test is consistently documented.

We would also expect that different default values for some Darwin Core terms may be useful for different communities. For example, WGS84 (or EPSG:4326) as a default for dwc:geodeticDatum may be a logical default in an international context as it remains the

default in most GNSS (Global Navigation Satellite System) receivers and smartphones, but a national geodeticDatum such as GDA94 in Australia may be a more acceptable default, or even be legislated in some circumstances.

We have flagged a subset of tests that will require the setting of a Parameter appropriate to the environment in which the tests are run, for example:

- VALIDATION_GEOGRAPHY_NOTSTANDARD
- VALIDATION_CLASSIFICATION_AMBIGUOUS
- AMENDMENT_YEAR_STANDARDIZED
- AMENDMENT_GEODETICDATUM_ASSUMEDDEFAULT

Moreover, spatial intersections will require some form of spatial buffering. For example, the test VALIDATION_COORDINATES_COUNTRYCODE_INCONSISTENT, without some form of spatial buffer, will assume high accuracy and precision on both the geographic coordinates and the available country boundaries. As this is unlikely, one needs to take terms like dwc:coordinateUncertaintyInMeters into account, or add a default spatial buffer to the coordinates and/or the country boundaries for the test to be meaningful in the majority of circumstances.

The Data Quality Framework requires that a Validation be related to an Information Element, but allows the Information Element to be specified in general terms (e.g., temporal information (see the Information Element Category in Table 1 and Suppl. material 4)) or in specific terms (e.g., dwc:eventDate). It is important to specify exactly which Darwin Core terms and other resources are required for each test. Thus Information Elements for the tests have been specified as particular Darwin Core terms. There are also specific scope descriptors for tests that identify if a test requires access to resources beyond available Darwin Core records. For example, the test AMENDMENT_GEODETICDATUM_STANDARDIZED requires a lookup table/vocabulary of values of possible geodetic datums for validation.

5.3 Test Types

We originally classified tests as having a severity of one of two states, either 'warning' or 'error', but discrimination between these states was recognised as being context dependent. We therefore decided to use the Framework's Data Quality Dimension terms (Veiga et al. 2017) as noted above to classify the nature of the issue that resulted in the flagged test. The terms suggested for expected status responses for **VALIDATION** tests are "Ambiguous", "Detected", "Empty", "Inconsistent", "Not found", "Not standard" and "Out of range" (see definitions of each in Suppl. material 1). "Ambiguous" is used where the interpretation of a Darwin Core term cannot be determined with certainty, e.g., a dwc:eventDate of "3/6/2017" could be either March 6, 2017 or June 3, 2017. The term "Detected" highlights where a Darwin Core term is "NOT_EMPTY" and we believe the user of the data needs to consider the implications, for example, VALIDATION_IDENTIFICATIONQUALIER_DETECTED highlights a potential taxonomic issue. "Empty" similarly highlights the situation where a Darwin Core term is either not present, or has no

value when we believe it is ideally needed, for example, "dwc:basisOfRecord is empty". "Inconsistent" is applied where there is a difference between two or more Darwin Core terms, for example, the supplied value for dwc:country does not match the supplied value for dwc:countryCode. "Not found" flags the situation where a Darwin Core term cannot be verified against an accepted source authority, for example, dwc:genus value is not found in the accepted source taxonomic authority. "Not standard" is used when the value of a Darwin Core term does not agree with the vocabulary in the accepted source authority or term specification, for example, "dwc:day is not standard" if it is not an integer in the range 1–31 inclusive. "Out of range" is used where ranges are known, for example, dwc:decimalLatitude is invalid if it is outside the range of -90 to 90 inclusive.

For **AMENDMENT**s, the responses are "Assumed default", "Converted", "Standardized" and "Transposed" (see definitions of each in Suppl. material 1). "Assumed default" occurs when a Darwin Core term is empty and we want to flag that an assumed value will be used, for example, if dwc:geodeticDatum is "EMPTY", it may be assumed to be for example, "WGS84" (EPSG:4326). "Converted" is used to flag that some form of coordinate conversion has been applied, for example, for AMENDMENT_ COORDINATES_CONVERTED, we may convert dwc:decimalLatitude and dwc: decimalLongitude with a dwc:geodeticDatum of "GDA94" to "WGS84(EPSG:4326)". Note that we would align with the broader community to always recommend that original values are never overwritten. "Standardized" is the response where we have considered the input values and have successfully converted one or more to a standard form, for example, "dwc:basisOfRecord standardized from the Darwin Core vocabulary of accepted values". Several amendments take the form "...from X" where X is a Darwin Core term, as in for example, "dwc:eventDate from dwc:verbatimEventDate".

MEASUREs were devised that summarised test results for each record (and can be accumulated across multiple records). Definitions of terms are in Suppl. material 1:

1. The number of validation tests that returned COMPLIANT,
2. The number of validation tests that returned NOT_COMPLIANT,
3. The number of validation tests that returned PREREQUISITESNOTMET,
4. The number of amendments PROPOSED and
5. The dwc:eventDate precision in seconds.

NOTIFICATIONs give only one response: 'NOTEMPTY' (see Suppl. material 1), for example, "dwc:dataGeneralizations NOTEMPTY" to flag to the user of the data that they need to consider the validity of the data for their purpose.

5.4 Implementation

We would strongly encourage the application of the core tests from data source to data use. For example, applying the tests as close as possible to the point of origin will short circuit the need for biodiversity data aggregators such as GBIF, the ALA and iDigBio to redirect records with issues resulting from the application of the tests back to the point of origin. In many cases, the 'point of origin' may no longer exist. Data capture tools such as iNaturalist, Project Noah, BioCollect, OzAtlas, etc. should ideally run the tests as data are entered into the application. Early detection of problems provides the most efficient way of addressing them. At the time of writing, GBIF, the ALA, and iDigBio have agreed to implement the core tests once they are finalized by the TDWG Data Quality Tests and Assertions Task Group. We also see the utility in making the tests available in a standard form via APIs, such as those used by GBIF (rGBIF: https://cran.r-project.org/web/packages/rgbif/index.html) and the ALA (ALA4R: https://cran.r-project.org/web/packages/ALA4R/index.html and http://api.ala.org.au).

Under some serial workflow forms of implementation, many of the tests have other tests as prerequisites, for example, "AMENDMENT_BASISOFRECORD_STANDARDIZED" has as a prerequisite the test "VALIDATION_BASISOFRECORD_NOTSTANDARD". We have documented the order of all prerequisites for implementers, should they choose to perform a single sequence of tests. However, the expected use of a set of tests is to run all VALIDATIONS and all MEASURES (potentially in parallel on a data set) to produce a pre-amendment view of the quality of the data, then to run all AMENDMENTS, and then to re-run all VALIDATIONS and all MEASURES to produce a post-amendment view of the quality of the data. Also, the order in which amendments to add data are run may be important. For example, if dwc:eventDate is empty, an attempt to populate it should firstly be from dwc:verbatimEventDate; if that is not possible, then from dwc:year, dwc:startDayOfYear and dwc:endDayOfYear, then if that is not possible from dwc:year, dwc:month and dwc:day. This is a particular caution to implementers who wish to run the tests in parallel.

6. Case Studies - Validating Date Fields using Core Tests and Assertions

While developing the test and assertion definitions, two case studies were undertaken using snapshots (GBIF 2019a and GBIF 2019c) of data from the GBIF occurrence data store. These case studies served to not only exercise the tests on large datasets, but also to show the results of a few of the tests on data before and after the GBIF interpretation process (GBIF 2019b). The case studies were performed with the event date-related tests, none of which require an external vocabulary. In one case, tests were run with the Java event_date_qc library (Morris 2019) on a workstation producing framework formatted results; in the other case, tests were run using Structured Query Language (SQL) on a copy of a snapshot loaded into Google BigQuery (Sato 2012).

Perhaps the most important outcome from the case studies was that the process of developing implementations provided feedback for both the test definitions and the development of the framework. Several rounds of redefinition and implementation were needed to make the tests complete, consistent, and able to be implemented in multiple technical scenarios.

6.1 Developing Date tests in the Kurator project

Prior to the formation of the TDWG Data Quality Interest Group, the FilteredPush Project (Morris et al. 2014) produced a data quality tool named FP-Akka (Morris et al. 2017). This tool included a large monolithic component for performing a set of tests on the value found in dwc:eventDate. This component concealed the complexity and linkages of this set of tests, and reported the results of these tests as a single data quality assertion. Over the course of the Kurator project (Morris et al. 2018), and in interactions with the Data Quality Interest Group and Tests and Assertions Task Group (TG2), this code was rewritten as a separate library of multiple small atomic tests, each following the (evolving) test definitions formulated by TG2. This library is the product: event_date_qc (Morris 2019 written in Java, available from the Maven central repository (Sonatype, Inc. 2011)). The process of developing the event_date_qc library involved multiple informative feedback loops with the process of defining standard tests in TG2.

The EventDate test actor (Morris et al. 2019) in FP-Akka made limited assertions (e.g., record is good, or record has a problem), but behind each assertion was a complex flow chart where many tests were performed to reach the conclusion that a record had a problem. Each test added to a long string of comments that accompanied the assertion, with the simple assertions masking the complexity underlying each. This comment set was provided to TG2, but proved difficult to compare with the other tests and was not included in the analysis above.

To approach the rewriting of this actor to provide tests consistent with the TG2 defintitions, we began by separating the actor out into multiple distinct tests, each developed from date-related issues observed in a few data sets (Harvard University Herbaria, Museum of Comparative Zoology, Southwest Collections of Arthropod Networks (SCAN), InvertEBase) and from test definitions developed in TG2, working with the Fitness for Use Framework (Veiga et al. 2017) (TG1) to develop test descriptions and responses in terms of the framework. We were thus led to a design that was test driven, with unit tests phrasing problems and expected solutions, and a division between low level library of methods containing test logic (e.g., is a text string in the form of an ISO date) separatated from the concerns of working with Darwin Core data inputs and reporting results in a manner consistent with the framework. This allows the library to provide low level tests independent of the framework, tests meeting the framework expectations, and to integrate into arbitrary data mapping and execution frameworks (such as within a Kurator actor within the Kurator workflow system (McPhillips et al. 2017)). The Kurator project also developed two libraries for working with framework concepts, and an OWL (Web Ontology Language) representation of the framework (Lowery et al. 2016 and Morris and Lowery 2018).

Key to the process of developing the test implementations and unit tests for these implementations were multiple feedback loops between TG2 discussions and code development, including running implementations of proposed test definitions on data. Similarly feedback loops examining the needs of reporting results to end users improved the TG1 framework development.

We encountered challenges along this path. These included changing the test definitions and framework, being moving targets for the implementation of test code, and the acquisition and development of suitable data for testing the implementation. In addition, understanding the framework and getting others to understand the framework proved a significant challenge. This was highlighted by repeated discussions within the task group of the nature of positive and negative assertions, the framework being designed around the positive concept of data having quality, while most of the participants in the task group were much more comfortable thinking about the negative sense of detection of errors in data.

Over the course of the feedback loops of test description and implementation, we settled on the following decisions:

1. Define small tests that are simple, not large tests embedding complex logic. Small tests can be mixed and matched in data quality profiles. Many small tests produce more assertions than large complex tests, but the simpler assertions are easier for data curators to understand and act upon. Small tests are much simpler for implementors to develop and maintain, and are much more likely to produce consistent results across different implementations.
2. Tests phrased as validations rather than issues. Most of the existing data quality tests that TG2 started examining were phrased in a negative sense, as issues, whereas the framework initially only allowed for the descriptions of data quality in a positive sense (see discussion under section 5.1).
3. Amendments paired with validations, as discussed in section 5.4.
4. Some tests must be able to take parameters to control the expected behavior when called for by differing use cases. We started out with a position that all implementors of a test should produce the same results on a given input data set, then realised that some tests are likely to use different authorities for different implementors, such as tests of scientific names, where the ALA is likely to want to test against an authoritative list of Australian taxa, while GBIF is likely to want to test against GBIF's backbone taxonomy. This led us to rapidly think about a subset of the tests needing to take parameters, and a guarantee of identical output for identical inputs - including the test parameters, vocabulary versions, and description of the inputs.

Deciding how to test whether or not some implementation of a set of tests produces the expected results is a challenge. For most of the tests, a unit testing approach is likely to be most effective, with unit tests examining the outputs of individual simple tests for a range of inputs, with the results (both values and status metadata) being known for each input. Such unit tests are relatively straightforward to implement for some tests, such as a

validation that tests to see if the value of dwc:day is in range (i.e., an integer in the range 1 to 31 inclusive). Other tests, such as an amendment that proposes a value for dwc:eventDate based on the value found in dwc:verbatimEventDate, are much less feasible to evaluate exhaustively for both expected good results and expected failure conditions. Unit tests by implementors would allow us to avoid the additional complexity of developing complete test data sets and files containing expected outputs to validate the behavior of implementations of test suites (where, among other issues, an implementation may not guarantee the output order of result rows from run to run). If entire test data sets are developed, then we note that it is important to be able to unambiguously mark test data as being synthetic, or being synthetically modified from actual data, to reduce the risk of synthetic test data sets being incorporated by mistake into scientific analyses.

Data quality test results could be produced in multiple forms, for quality control purposes; they could be presented as detailed reports on data sets for consumption by the curators of the database of record for those data sets, or they could be attached as annotations to copies of the data. In anticipation of this use, we developed an RDF representation of the framework, and of data quality results from the framework, with the expectation that data quality assertions could be wrapped in W3C annotations. In data quality control reports, it is likely that the subset of records of most interest are those for which validations are not compliant pre-amendment, but are compliant post-amendment (that is, the records for which amendments propose changes that improve the quality of the data for the core fitness purposes). For quality assurance (QA) purposes, reports could potentially have much simpler summaries of measures of the percent of records in the data set (100% for QA) that comply with the validations that test if the data are fit for the use.

Over the course of the rounds of developing test definitions and implementing tests, we developed a set of tools to support the coding and test result production including annotations for the Java code implementation of the tests that link particular Java class methods to particular tests, an RDF representation of the framework, and a tool to generate RDF descriptors and stub Java code of tests from a CSV representation of test descriptions (Table 1).

6.2 Java Case Study

The Java event_date_qc library (Morris 2019), implements all of the time-related tests from the core test set (as well as a few others that are no longer core). These were run on a data set of all specimen-based occurrence records from GBIF as of August 2019 (GBIF 2019a - 170,724,036 occurrence records where dwc:basisOfRecord is dwc: PreservedSpecimen or dwc:basisOfRecord is dwc:FossilSpecimen, https://doi.org/10.15468/dl.bwcpqx). The results of this run are shown in Table 2.

Table 2.

Results of a run of event_date_qc on about 170 million specimen-based occurrence records from GBIF (GBIF 2019a).

Result Status	Result Value	Pre-Amendment	Post Amendment	Percent Change
Test: VALIDATION_EVENT_EMPTY				
HAS_RESULT	COMPLIANT	133,438,715	133,438,715	0.00%
HAS_RESULT	NOT_COMPLIANT	36,455,664	36,455,664	0.00%
Test: VALIDATION_EVENTDATE_EMPTY				
HAS_RESULT	COMPLIANT	93,764,511	130,667,642	21.72%
HAS_RESULT	NOT_COMPLIANT	76,129,868	39,226,737	-21.72%
Test: VALIDATION_EVENTDATE_EMPTY				
DATA_PREREQUISITES_NOT_MET	--	76,129,868	392,26,737	-21.72%
HAS_RESULT	COMPLIANT	74,990,834	123,912,003	28.80%
HAS_RESULT	NOT_COMPLIANT	18,773,677	6,755,639	-7.07%
Test: VALIDATION_EVENTDATE_OUTOFRANGE				
DATA_PREREQUISITES_NOT_MET	--	94,903,545	45,982,376	-28.80%
HAS_RESULT	COMPLIANT	74,492,269	123,342,156	28.75%
HAS_RESULT	NOT_COMPLIANT	498,565	569,847	0.04%
Test: VALIDATION_EVENT_INCONSISTENT				
DATA_PREREQUISITES_NOT_MET	--	110,520,256	44,225,051	-39.02%
HAS_RESULT	COMPLIANT	47,957,506	116,519,824	40.36%
HAS_RESULT	NOT_COMPLIANT	11,416,617	9,149,504	-1.33%
Test: VALIDATION_EVENT_INCONSISTENT				
DATA_PREREQUISITES_NOT_MET	--	81,675,691	43,361,643	-22.55%
HAS_RESULT	COMPLIANT	86,848,963	125,637,244	22.83%
HAS_RESULT	NOT_COMPLIANT	1,369,725	895,492	-0.28%
Test: VALIDATION_YEAR_EMPTY				
HAS_RESULT	COMPLIANT	93,300,36	126,852,070	19.75%
HAS_RESULT	NOT_COMPLIANT	76,594,013	43,042,309	-19.75%
Test: VALIDATION_YEAR_OUTOFRANGE				
DATA_PREREQUISITES_NOT_MET	--	77,056,465	43,504,761	-19.75%

Result Status	Result Value	Pre-Amendment	Post Amendment	Percent Change
HAS_RESULT	COMPLIANT	92,495,590	125,534,971	19.45%
HAS_RESULT	NOT_COMPLIANT	342,324	854,647	0.30%
Test: VALIDATION_DAY_NOTSTANDARD				
DATA_PREREQUISITES_NOT_MET	--	87,302,430	43,502,938	-25.78%
HAS_RESULT	COMPLIANT	81,180,819	124,980,313	25.78%
HAS_RESULT	NOT_COMPLIANT	1,411,130	1,411,128	0.00%

We draw attention to a few of the results shown in Table 2. VALIDATION_EVENT_EMPTY, which tests for some value in at least one of the dwc:Event terms, shows a 0% percent change pre- and post-amendment, as there are no amendments that are able to propose values from elsewhere in Darwin Core, if there are no values in any of the dwc:Event terms, then none are inferred. There are other possible amendments that could propose values for an event date from other data, such as inferring a possible range of dates collected from a collector name (Dou et al. 2012). In contrast, the results of VALIDATION_EVENTDATE_EMPTY, which tests whether dwc:eventDate contains a value, has an increase of 21.7% compliance pre- and post-amendment, as in many cases values for dwc:eventDate could be inferred from values in dwc:day, dwc:month, dwc:year, dwc:startDayOfYear, dwc:endDayOfYear, and dwc:verbatimEventDate. This increase is mirrored in VALIDATION_EVENTDATE_NOTSTANDARD, where the cases of prerequisites not met (no value present in dwc:eventDate) pre- and post-amendment decrease 21.7%, while the compliant cases increase by some 28%, adding the populated dwc:eventDate values to ones that have been standardized by other amendments.

6.3 SQL Case Study

A snapshot of the entirety of the GBIF occurrence data store as of 15 April 2019 (1,213,409,995 occurrence records https://doi.org/10.15468/dl.5pmzev) (GBIF 2019c) was exported by GBIF as a set of Apache Avro files, which were loaded into a table in Google BigQuery. From this table, distinct combinations of the date-related dwc:Event terms (v_eventdate, v_year, v_month, v_day, v_verbatimeventdate, year, month, day, v_startdayofyear, v_enddayofyear) were extracted into a new table ('gbif_dates') along with the number of corresponding occurrence records ('occcount'). This table was queried using SQL statements (see Suppl. material 3) constructed to extract the number of occurrence records matching the expected responses for a set of 9 event date-related validation tests. The tests run were: EVENT_EMPTY, EVENTDATE_EMPTY, EVENTDATE_NOTSTANDARD, EVENTDATE_OUTOFRANGE, EVENTDATE_INCONSISTENT, YEAR_EMPTY, YEAR_NOTSTANDARD, MONTH_NOTSTANDARD and DAY_NOTSTANDARD. Two dwc:eventDate tests (TG2_VALIDATION_EVENTDATE_OUTOFRANGE and TG2_VALIDATION_EVENTDATE_INCONSISTENT) were omitted due to the complexity of implementing them using SQL queries. The GBIF data processing pipeline

(GBIF April 2019) produces three event date-related fields (year, month, day) in addition to those acquired from the original sources. For these three fields we ran the tests appropriate to those terms for both the original and interpreted data (Suppl. material 3).

For each test we obtained counts of occurrence records for the categories of expected response (INTERNAL_PREREQUISITES_NOT_MET, NOT_COMPLIANT, COMPLIANT) from which we calculated the percent of occurrence records in the snapshot (Table 3). Corresponding numbers are found in the Suppl. material 3. In some cases, it was not possible to distinguish between results from the categories INTERNAL_ PREREQUISITES_NOT_MET and NOT_COMPLIANT, because in the GBIF snapshot we did not have access to the original data, so we were unable to determine if an EMPTY value was due to a missing value or a missing field. In these cases the results for the two categories are combined. The tests for year, month and day were run against the values from the source (v_year, v_month, v_day) as well as against the GBIF-interpreted values (year, month, and day). The significant decrease in percentages in the category INTERNAL_ PREREQUISITES_NOT_MET between the corresponding tests is due to the fact that GBIF processing is able to interpret the verbatim event fields and provide year, month and day when they were not given explicitly in v_year, v_month, and v_day. The results from the GBIF-interpreted values (year, month, day) are in parentheses in Table 3. The full results are given in Suppl. material 3.

Table 3.

Percentages of occurrence records in 2019-04-15 snapshot of GBIF-mediated data (GBIF 2019c) that fit the three categories of expected responses for each of the event date-related validation tests run. Results of the tests that were run against the GBIF interpreted versions of the data are included in parentheses. Counts can be found in (Suppl. material 3). (NA = not applicable; ND = not determined)

TG2_VALIDATION Test	INTERNAL_PREREQUISITES_NOT_MET	NOT_COMPLIANT	COMPLIANT
EVENT_EMPTY	NA	5.8%	94.2%
EVENTDATE_EMPTY	NA	55.4%	44.6%
EVENTDATE_NOTSTANDARD	55.4%	3.6%	41.4%
EVENTDATE_OUTOFRANGE	55.4%	ND	ND
EVENTDATE_INCONSISTENT	77.1%	ND	ND
YEAR_EMPTY	NA	28.3% (7.6%)	71.7% (92.4%)
YEAR_NOTSTANDARD	28.3% (7.6%)	0.2% (0%)	71.5% (92.4%)
MONTH_NOTSTANDARD	30.5% (9.9%)	0.2% (0%)	69.3% (90.1%)
DAY_NOTSTANDARD	31.7% (10.8%)	0.1% (0%)	68.1% (89.2%)

In this case study we have applied event date-related validation tests against a large body of aggregated biodiversity occurrence data. The purpose was not to explore in-depth the underlying causes for the state of data quality found, but rather to show that data quality validation tests can be applied even at the largest of scales via an alternative

implementation. Nevertheless, it is clear from the before and after year month and day tests that GBIF's data aggregation pipeline has a significant positive effect on improving the quality of the data for these concepts.

7. Developing Controlled Vocabularies of Values

Over many years of working with biodiversity data, the community has repeatedly come to the conclusion that it is of pressing importance to have shared vocabularies. This conclusion also arises from the work of building the Tests and Assertions, where 32 of the core tests have been identified as needing a controlled vocabulary. To put it in general terms, in any process of sharing information, the sender and the receiver of the information need to use common codes and signifiers for it to be of any use to the receiver. The representation of this relationship (proposed by Ogden and Richards 1923) is a triangle, composed of a symbol, a thought and a referent (Fig. 15). The symbol can be pictured as a word (e.g., “tree”), the thought would be the ideal representation of the word (e.g., our mental representation of a *tree*), and the referent would be the real object to which both the symbol and the thought are related (e.g., the tree itself as the object). In this context, we can have full vocabularies that are accumulations of symbols, sets of words that describe thoughts and have referents. Vocabularies can be restricted to certain portions of our universe and apply at different scales of our description of the world. For instance, we can have vocabularies to describe geographical administrative divisions (“continent”, “country”, “municipality”, etc.) and other, more specific vocabularies to describe elements within those administrative divisions (“Africa”, “Asia”; “Nigeria”, “Indonesia”, etc.). There are cases in which the triangle symbol-thought-referent may turn into more complex shapes (Fig. 15). Among these cases are the polysemy, homonymy and synonymy. Polysemy happens when more than one thought and referent are related to a symbol (i.e., the same word, having the same etymology - that is, the same origin, describes more than one concept). An example of it is the word “wood”, with a single origin in Germanic language, wudu in old English, and that can be applied to a piece of a tree or to an area with many trees (Fig. 15B). Homonymy also happens when there is more than one thought and referent related to a symbol, but the symbols have different etymologies (i.e., the same word, with different etymology, describes distinct concepts). An example of it is the word “bank”, from Old Norse “bakki” to describe the side of a river or lake, from Latin “bancus” to describe a financial institution (Fig. 15D). Finally, synonymy happens when one thought and referent are described by more than one symbol (i.e., a single concept represented by different words). An example of this case would be the words “freedom” and “liberty”, both referring to the quality or state of being free (Fig. 15C).

The sharing of biodiversity information is not free of the complexities described above. We face the same problem of the receiver understanding the sender’s message, effectively capturing its meaning, despite homonymy, polysemy and synonymy. Furthermore, understanding the meaning sometimes requires expert knowledge. For communication to be effective, we need to decide on common vocabularies that unequivocally refer us to the same concepts.

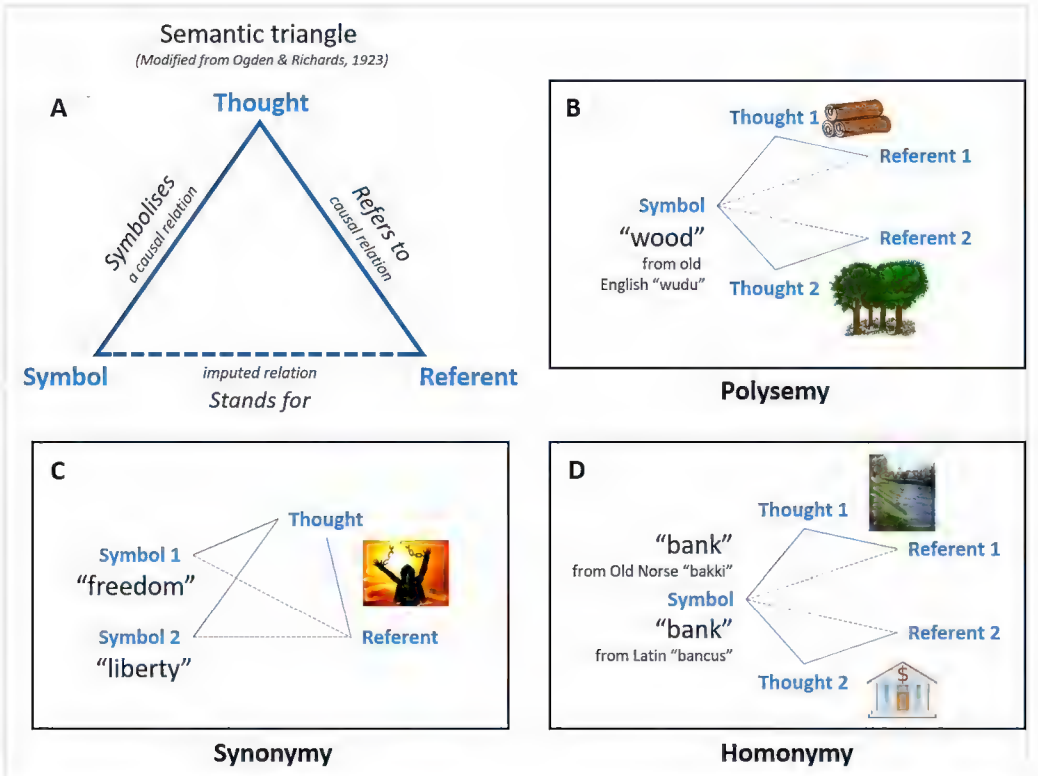

Figure 15.

(A) Semantics triangle and exemplary cases of (B) polysemy, synonymy (C) and (D) homonymy.

The biodiversity data community has agreed upon some vocabularies for sharing information, i.e., biodiversity data standards for sharing information, among which the Darwin Core standard (Wieczorek et al. 2012) is the most widely used. This standard is a set of terms and definitions related to biodiversity data, and its terms can be thought of as the names of the columns in a spreadsheet where we are capturing data. For example, there are terms such as dwc:catalogNumber, dwc:genus, dwc:stateProvince*1. This allows us to capture and share certain information that we agree belongs under one of those terms. For example, we agree that if we have a record of an organism that is a *female*, we will share the fact that it is a female under the "sex" term (dwc:sex term in Darwin Core). However, we have not yet reached an agreement about how to express the value *per se*. For instance, we could represent *female* with the words "female", "fem.", "f.", and all the possible variants that derive from the different forms of abbreviation of the word in combination with the language (e.g., in Spanish it would be "hembra", or "h.", and so forth). For relatively simple concepts, such as the sex of an individual, one may think that the possible variants are finite, and that they would only be referring to a handful of concepts, to which most people can easily relate (in the example, the concept *female*). For some other concepts there may be variation derived from the use of the same words in certain languages to describe different concepts (homonymy or polysemy) (see Fig. 15). We could think of one kind of behaviour as an example. If the information is that an animal was taking in food, we may capture it under the term behavior in different ways: "feeding", "eating". "Eating" raises no doubt about the subject receiving the food, but "feeding" may well mean that the animal is feeding itself or others. Therefore, in the latter case we are using the same word to describe different behaviors, ultimately different concepts (polysemy). The reverse, using different words to describe the same concept (synonymy), is also common. In turn, the worst case appears when different people accept different

definitions of a particular concept. This is often bound to expert knowledge needing to define and understand a concept. One of the most intricate examples of complex concepts are biological taxonomies, encompassing how we name distinct species and species concepts.

With all the above in mind, it appears evident that faithful communication of a piece of biodiversity information among people is not trivial, and computers can add to the complexity. People working with biodiversity data have already learnt to discern concepts and words, but computers still need to undergo that learning process. For example, an English speaking person can confidently interpret "f." as "female" meaning *female* if they see it written under the sex term, but a computer needs to be told in advance that "f." is a variant of "female" and that "female" means *female*. This would allow it to show the variants as representing the same concept and to relate them to other concepts (such as, for example, "has gynoecium" if it is an angiosperm plant). In this sense, semantics and ontologies play a central role in providing strict word/concept definitions and establishing the relationships between them, which can be used by computers. Several ontologies are being developed that refer to biological concepts, such as the description of organisms in collections (e.g., Biological Collections Ontology, BCO), and the description of ecological processes (e.g., Environment Ontology, ENVO, Buttigieg et al. 2013). Although much work has been done in this respect, we currently do not possess a full suite of vocabularies to apply uniformly across the biodiversity data community.

It is worth asking why we consider the lack of vocabularies to be a data quality issue and what the consequences are of not having common vocabularies. Firstly, heterogeneity in the data (i.e., presence of variants for certain values) renders data less discoverable and very difficult to use. This heterogeneity works against application of the FAIR data principles (Wilkinson et al. 2016), particularly those that expect data to be findable and reusable, but also by extension affecting accessibility and interoperability when using the data. Revisiting the sex example, if a researcher were to filter a database using the word "female", records that refer to *females* but that have "f." as a sex value would not be returned. Also, even if they retrieved all possible records to avoid that gap, they would currently have to manually determine what "f." (or "h." in Spanish) stands for. Without common vocabularies, and the capture of information in myriad ways, we risk being incomplete and inaccurate in our transmission of information. If we cannot be certain that a particular value unambiguously refers to a particular concept, we cannot assert that a record containing such value could reliably be used for a particular purpose. It is in this context that, where possible, the construction and use of vocabularies of values, including an explicit declaration of usage, represents a matter of data quality.

There are already many vocabularies of values used by the biodiversity data community. However, to date, those vocabularies are only used by certain portions of the community, there is no general consensus, and there is no single repository enabling the exploration of the available resources. While some of the available vocabularies are discipline-specific (e.g., vocabulary to describe life stages in marine organisms), many that could be applied more broadly remain independent and scattered. Additionally, similar lists of terms that refer to the same concepts can be found in different languages, but disconnected from one

another. Several reasons may account for this current state of vocabularies, spanning from economical limitations (lack of resources to allocate to merging initiatives) to social/ personal impediments (where certain sectors of the community are disinclined to adopt new community-driven practices). Irrespective of these causes, there is a growing acknowledgement of the need for consensus.

The TDWG Data Quality Interest Group has begun to tackle this problem, with the aim of creating a suitable environment for thought and development of **vocabularies of values**. Accordingly, a new task group has been established to: (1) prepare a scoping document to determine the types of vocabularies needed (including multi-lingual approaches) and the strategy for organizing the construction and/or management of new/existing vocabularies; (2) develop a common repository to store vocabularies and/or link to existing ones; (3) develop a standard format for building TDWG vocabularies; and (4) develop an exemplary vocabulary following a standard format. These actions aim at providing the community with a framework to work and build upon vocabularies of values to foster better understanding and interoperability. As a data quality problem, the availability and use of vocabularies of values is necessary for testing and asserting the fitness of biodiversity data for particular purposes, which are central to the activities of the Data Quality Interest Group and of TDWG as a whole.

8. Discussion: global implications and future directions

In this paper we present a data quality framework and a series of use cases, and show how the framework can be used to make consistent descriptions of data quality tests, how the tests can be applied to assess and enhance real world data through an example implementation, and how the results can be phrased in the vocabulary of the framework. The aim has been to gain more consistency between data publishers in relation to evaluation of data fitness for use, to make the data more trustworthy, and to maximise the potential re-use of data. The approach is to target data providers, data users, and data aggregators.

Data producers may, in the majority of cases, be the best suited to determine the quality of their own records (e.g., they know the exact location where they collected a specimen; the date on which they made an observation). Unfortunately, in the majority of cases data producers may not be available to perform data quality checks. In reality, the closest to the source that one may reasonably get are the data providers (which may or may not be the primary data producers). Being closest to the source, they uniquely possess knowledge beyond the record level that can inform data quality evaluations and may be able to improve the data quality. This makes data providers a key target of the work presented here. Although it is expected that aggregators will have greater capacity than individual data providers to perform certain operations on the data records, there is still information currently only available from data providers. A classic example comes from natural history collections, where errors in coordinates arise from misinterpretations of the collector's handwriting, and which might be known to curators (e.g., substitution of a "1" for a "7"). Even where aggregators have full capabilities for dealing with all data quality aspects,

repatriation of corrected/enhanced data to the data providers remains a broadly unsolved challenge. These arguments reinforce the idea that improving data quality as close to the source as possible is highly recommended (Chapman 2005a, Belbin et al. 2013).

For data users, the proposed framework, tests and profiles offer the possibility of choosing which data to use according to their data quality needs and adjusting the tests and profiles in accordance. It is likely that a large proportion of users would demand similar data quality checks (e.g., missing or mismatching coordinates, missing or ambiguous dates, unrecognised scientific names). In this sense, it would seem most appropriate, or at least most efficient, that aggregators are the ones tasked with the implementation of the tests, reporting on the quality of the data, and allowing for searches using specific quality filters. This would be especially true for assessing presence/absence of values, levels of precision and levels of confidence in a consistent way. However, for those researchers who require different or more exhaustive data quality checks, the framework, tests and profiles still offer the standardised foundation upon which to base a custom implementation. As more and different types of data become available, it is possible that the kinds of questions that are asked will become more specific and complex. In this sense, it is fundamental to enable users to adjust at their own pace, while aggregators make these capabilities broadly available through large-scale processing.

Data aggregators play a fundamental role in exposing data more widely, assessing and where possible, improving the 'quality' of data. The work presented here describes the foundations for standardised data quality practices that can be replicated across data publishers, and in this case, across aggregators. To aggregators, this work provides a framework with covenanted terminology that should improve interoperability. Also, it provides definitions for a core set of tests that the community has highlighted as essentially relevant for the most common uses, therefore setting, from the bottom-up, the minimum data quality tests that the community would expect to have implemented. The profiles, in turn, are also those most commonly wanted by the community, and provide an indication of how we could offer data that is more likely to be suited for particular uses. It is expected that aggregators would adopt these standardisations and make them part of their workflows, so that, independently of the implementation processes they each run, the users can be assured that they are getting data that has undergone conceptually the same basic analyses and modifications, regardless of the aggregator consulted.

The implementation of the core tests and assertions by aggregators will greatly help to align 'data quality' reporting. Data providers, however, may not have the resources to either react to the increasing numbers of reports or to implement the tests themselves, or may not have the processes in place to repatriate the data into their own databases. We would hope, however that data providers would see the considerable ongoing cost-benefits of incorporating the tests into their infrastructure over the long term.

A key to solving data quality issues is the actual implementation of the concepts presented here. Developing a single standardised implementation is not possible in the short term because of the diversity of infrastructures among aggregators. It is clear, however, that a natural next step would be to standardise and thoroughly document a core set of

expectations for the behavior of implementations of standard test definitions, including specifications of expected outputs for particular inputs. Such a set of explicit expectations for the behavior of test implementations would allow tests to be independently implemented across aggregators, yet have different implementors make consistent assertions. This would include a consistent methodology for documenting and reporting data quality assertions, including in reports and in annotations. In the near future, it might be a responsibility of the aggregators, and an overall community effort to do data quality processing in a centralised index. Improved data could later be re-ingested by any given aggregator to display in any way they deem appropriate. Having only a set of consistent core test definitions, and a description of the expected response structure from a test, we are some steps away from that ideal scenario, but important progress has been made towards that purpose. For instance, in 2018, during the second Global Biodiversity Informatics Conference (GBIC2, Hobern et al. 2019), the community agreed upon the importance of taking collaborative approaches to design, fund, implement and sustain infrastructure components and tools required by multiple stakeholders, and much was discussed about centralising processes that are broadly utilised.

Moving forward it will become relevant to review whether the use of the Darwin Core standard remains an appropriate and sufficient way to share biodiversity data, and to complement it or replace it as appropriate. Also, it will be of special interest to focus on precision and uncertainty in the resolution of data quality issues. Over time, and as more data become available and are semantically interconnected, it will become more important to evaluate and declare the references, particularly linked open data used in data quality resolutions, and to provide confidence levels associated with a given result. For instance, it should be possible to answer questions such as: what is being referenced, how precise is the determination, what is the level of confidence in this determination, and what evidence feeds into the assessment.

This initiative provides a necessary baseline. Ideally, the data quality issue would be approached from multiple perspectives, providing all stakeholders with tools that respond to their different needs and degrees of expertise. Such actions could include: (a) fostering the completion and continuous improvement of resources such as the Catalogue of Life (Catalogue of Life 2019); (b) unifying data indexing as a shared enterprise delivering persistently identified records; (c) reporting directly inside repositories on missing elements that would improve the interpretability of data; (d) facilitating metadata-level defaulting of values not contained in records; (e) supporting post-publication completion/amendment of data records by community consensus; and (f) providing data publication tools that allow researchers to publish exactly what they have in exactly the structure they have it, with clear reporting of how the data will be interpreted as Darwin Core records, with tools for them to adjust this interpretation. The impact of any of these individual actions will be dependent on the extent to which community amalgamation is attained.

It is hoped and anticipated that the processes outlined within this paper will contribute substantially, and lead to increased consistency at all levels, from local and regional collections institutions, national initiatives such as the ALA and iDigBio, domain-specific initiatives such as VertNet, and globally through GBIF and OBIS, etc. The eventual

incorporation of the tests into software systems such as iNaturalist, Specify, Symbiota, Brahms, and others would go a long way toward having the tests conducted close to the source, which would be an ideal outcome. Some citizen science initiatives such as iNaturalist and eBird, have expressed an interest in incorporating the tests into their data quality control systems, thus expanding the concepts to observation data, beyond physical specimens. While the community is focusing on finding the best ways to collaborate, the work presented here sets a broad conceptual schema around which a formalisation can be built that will assist in achieving the higher integrated goal.

9. Conclusion

There is a great need for a framework and standards to address the quality of the billions of biodiversity data records being openly shared. The process of developing those standards to cater to all levels of the data chain—from collection, through curation and storage, publication, aggregation and finally to the end users—is complex. The TDWG Data Quality Interest Group has addressed this process by developing an overall fitness for use framework, extracting and studying use cases, developing a core set of tests and assertions, and beginning the process of developing vocabularies of values. The progress presented in this paper represents many person-years of effort, but much work remains to be done. For example, we need to develop test datasets, plan and organise for the development of essential vocabularies of values, and bring together the many stakeholders and custodians to ensure all data records have maximum potential for re-use. The Task Group on Tests and Assertions has concentrated on the core set of tests, but additionally there are hundreds of tests that were set aside as not being essential across domains, being too complex to implement widely, or not sufficiently powerful or discriminating. To make all this work, we need to develop efficient feedback mechanisms. We hope that the use of standard annotations (being developed by the TDWG Annotations Interest Group) will greatly help with this. We also recognise the significance of Globally Unique Identifiers (GUIDs) at the record level for feedback mechanisms to work best, so we would urge institutions and others to begin this process as soon as practical.

Acknowledgements

Many people have been involved and contributed to the discussion and work of the TDWG Data Quality Interest Group. We have over 100 people who have listed an interest in the Group, and while not all have been actively involved, we value the contributions of those that have. We would particularly like to thank Global Biodiversity Informatioin Facility (GBIF), the Atlas of Living Australia (ALA), iDigBio, the Kurator Project, VertNet, the University of São Paulo, and the Biodiversity Information Standards (TDWG) Executive for supporting travel for the very-necessary face-to-face meetings between TDWG Conferences. We would also like to acknowledge support for the publication from Pensoft as part of their Silver Sponsorship at the 2016 TDWG Annual Conference in Santa Clara de San Carlos, Costa Rica.

Funding program

Travel of participants to various meetings was supported by the Atlas of Living Australia, US National Science Foundation Division of Biological Infrastructure, Fundação de Amparo à Pesquisa do Estado de São Paulo, the University of São Paulo, iDigBio, VertNet, the Global Biodiversity Information Facility and the Biodiversity Information Standards (TDWG) Community Support Fund.

Several organisations provided financial support for some of the time for the project, including, The Atlas of Living Australia for Lee Belbin, Miles Nicholls and Emily Rees, the US National Science Foundation Division of Biological Infrastructure for Paul J. Morris and John Wieczorek, the University of São Paulo and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq grant # 308326/2010-5 and 311531/2014-8), for Antonio M. Saraiva, the University of São Paulo and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES grant # 233676/2014-7) for Allan Koch Veiga, and iDigBio for Alex Thompson.

Support for meetings in São Paulo, Brazil, Canberra, Australia, Monash, Australia, and Gainesville, USA was provided by Fundação de Amparo à Pesquisa do Estado de São Paulo, the University of São Paulo, the Atlas of Living Australia, the Global Biodiversity Information Facility and iDigBio.

Grant title

Kurator: A Provenance-enabled Workflow Platform and Toolkit to Curate Biodiversity Data. **NSF:DBI**:1356438 and 1356751.

ABI Development: Collaborative Research: VertNet, a New Model for Biodiversity Networks **NSF:DBI**:1062193.

Biodiversity data quality: developing a common framework to improve fitness for use of biodiversity data. **FAPESP** grant #15/24168-3

Author contributions

Major areas of contribution are as follows; Framework: Allan Veiga, Paul Morris, Antonio Saraiva, Christian Gendreau, Arthur Chapman. Use Cases: Miles Nicholls, Emily Rose Rees, Dmitry Schigel. Tests and Assertions: Lee Belbin, Arthur Chapman, John Wieczorek, Paula Zermoglio, Paul Morris, Alex Thompson. Case Studies: John Wieczorek, Paul Morris. Vocabularies: Paula Zermoglio, Arthur Chapman, John Wieczorek. Editing and consistency: Shelley James, Abigail Benson. Overall coordination: Arthur Chapman.

References

- Anderson RP, Araújo M, Guisan A, Lobo JM, Martínez-Meyer E, Peterson AT, Soberón J (2015) Final Report of the Task Group on GBIF Data Fitness for Use in Distribution Modelling Are species occurrence data in global online repositories fit for modeling species distributions? The case of the Global Biodiversity Information Facility (GBIF). Unpublished https://doi.org/10.13140/RG.2.2.27191.93608
- Arnaud E, Castañeda-Álvarez NP, Cossi JG, Endresen D, Jahanshiri E, Vigouroux Y (2017) Final Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity. GBIF, Copenhagen. URL: https://www.gbif.org/document/82283/
- Beck J, Böllera M, Erhardt A, Schwanghart W (2014) Spatial bias in the GBIF database and its effect on modeling species' geographic distributions. Ecological Informatics 19: 10-15. https://doi.org/10.1016/j.ecoinf.2013.11.002
- Belbin L, Daly J, Hirsch T, Hobern D, LaSalle J (2013) A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. ZooKeys 305: 67-76. https://doi.org/10.3897/zookeys.305.5438
- Biodiversity Information Standards (TDWG) (2018) Darwin Core quick reference guide. http://dwc.tdwg.org/terms. Accessed on: 2019-11-26.
- Busby JR (1979) Australian Biotaxonomic Information System. Introduction and Data Interchange Standards. Australian Biological Resources Study, Australia: Canberra, 25 pp.
- Buttigieg P, Morrison N, Smith B, Mungall CJ, Lewis SE, the ENVO Consortium (2013) The environment ontology: contextualising biological and biomedical entities. Journal of Biomedical Semantics 4 (1). https://doi.org/10.1186/2041-1480-4-43
- Catalogue of Life (2019) Catalogue of Life. https://www.catalogueoflife.org/. Accessed on: 2019-10-05.
- Chapman A, Busby J (1994) Linking plant species information to continental biodiversity inventory, climate modeling and environmental monitoring. In: Miller RI (Ed.) Mapping the Diversity of Nature. Chapman & Hall, London, 218 pp. https://doi.org/10.1007/978-94-011-0719-8_11
- Chapman A (2005a) Principles of Data Quality, version 1.0. GBIF Secretariat, Copenhagen, 61 pp. [In English]. URL: http://www.gbif.org/document/80509 [ISBN ISBN 87-92020-03-8]
- Chapman A (2005b) Principles and Methods of Data Cleaning - Primary Species and Species-Occurrence Data, version 1.0. GBIF Secretariat, Copenhagen. [In English]. URL: http://www.gbif.org/document/80528
- Chapman A (2005c) Uses of Primary Species-Occurrence Data version 1.0. Report for the Global Biodiversity Information Facility. GBIF, Copenhagen, 100 pp. URL: https://www.gbif.org/document/80545
- Chrisman NR (1991) The Error Component in Spatial Data. Maguire D.J. et al. (eds) Geographical Information Systems 1: 165-174.
- Darwin Core and RDF/OWL Task Groups (2015) Darwin Core RDF Guide (URI: http://rs.tdwg.org/dwc/terms/guides/rdf/). Biodiversity Information Standards (TDWG) URL: https://dwc.tdwg.org/rdf/

- Dou L, Cao G, Morris PJ, Morris RA, Ludäscher B, Macklin JA, Hanken J (2012) Kurator: A Kepler Package for Data Curation Workflows. Procedia Computer Science 9: 1614-1619. https://doi.org/10.1016/j.procs.2012.04.177
- Edwards JL (2004) Research and societal benefits of the Global Biodiversity Information Facility. Bioscience 54: 485-486.
- Gaiji S, Chavan V, Ariño A, Otegui J, Hobern D, Sood R, Robles E (2013) Content assessment of the primary biodiversity data published through GBIF network: Status, challenges and potentials. Biodiversity Informatics 8 (2). https://doi.org/10.17161/bi.v8i2.4124
- GBIF (2018) Big data for biodiversity: GBIF.org surpasses 1 billion species occurrences. https://www.gbif.org/news/5BesWzmwqQ4U84suqWyOQy/big-data-for-biodiversity-gbiforg-surpasses-1-billion-species-occurrences. Accessed on: 2018-9-26.
- GBIF (2019a) GBIF Occurrence Download. GBIF.org (26 August 2019). https://doi.org/10.15468/dl.bwcpqx
- GBIF (2019b) GBIF Infrastructure: Data processing from publication to discovery. https://www.gbif.org/en/article/5i3CQEZ6DuWiycgMaaakCo/gbif-infrastructure-data-processing. Accessed on: 2019-10-24.
- GBIF (2019c) GBIF Occurrence Download. GBIF.org (15 April 2019). https://doi.org/10.15468/dl.5pmzev
- GBIF (2020) Free and Open Access to Biological Data. https://www.gbif.org/. Accessed on: 2020-1-12.
- Graham C, Ferrier S, Huettman F, Moritz C, Peterson A (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19: 497-503. https://doi.org/10.1016/j.tree.2004.07.006
- Hamilton A (2013) Evolution of Phylogenetic Systematics. University of California Press, 311 pp. https://doi.org/10.1525/california/9780520276581.001.0001
- Hobern D, Baptiste B, Copas K, Guralnick R, Hahn A, van Huis E, Kim E, McGeoch M, Naicker I, Navarro L, Noesgaard D, Price M, Rodrigues A, Schigel D, Sheffield CA, Wieczorek J (2019) Connecting data and expertise: a new alliance for biodiversity knowledge. Biodiversity Data Journal 7: e33679. https://doi.org/10.3897/BDJ.7.e33679
- Landuyt WV, Vanhecke L, Brosens D (2012) Florabank1: a grid-based database on vascular plant distribution in the northern part of Belgium (Flanders and the Brussels Capital region). PhytoKeys 12: 59-67. https://doi.org/10.3897/phytokeys.12.2849
- Lane M (2005) The Global Biodiversity Information Facility. Bulletin of the American Society for Information Science and Technology 30 (1): 22-24. https://doi.org/10.1002/bult.301
- Levatich T, Padilla F (2016) EOD - eBird Observation Dataset. Cornell Lab of Ornithology. GBIF https://doi.org/10.15468/aomfnb
- Longmore R (1989) Atlas of Elapid Snakes of Australia. Revised Edxition. Australian Flora and Fauna Series 7.
- Lowery D, Morris P, Veiga AK (2016) Kurator-Org/Kurator-Ffdq: Initial Release Of Kurator-Ffdq Library Version 1.0.0. Zenodo https://doi.org/10.5281/ZENODO.192186
- Mackay K (2017) Marine biological observation data from coastal and offshore surveys around New Zealand. Version 1.5. The National Institute of Water and Atmospheric Research (NIWA) https://doi.org/10.15468/pzpgop
- Maldonado C, Molina C, Zizka A, Persson C, Taylor C, Albán J, Chilquillo E, Rønsted N, Antonelli A (2015) Estimating species diversity and distribution in the era of Big Data: to

what extent can we trust public databases? Global Ecology and Biogeography 24 (8): 973-984. https://doi.org/10.1111/geb.12326

- McGeoch M, Groom Q, Pagad S, Petrosyan V, Ruiz G, Wilson J (2016) Data fitness for use in research on alien and invasive species. Global Biodiversity Information Favility, Copenhagen. URL: https://www.gbif.org/document/82958
- McPhillips T, Morris P, Lowery D, Zhang Q, Wieczorek J, Veiga AK (2017) Kurator-Org/Kurator-Akka: Kurator-Akka Version 1.0.1. Zenodo https://doi.org/10.5281/ZENODO.1068311
- Mesibov R (2013) A specialist's audit of aggregated occurrence records. ZooKeys 293: 1-18. https://doi.org/10.3897/zookeys.293.5111
- Mesibov R (2018) An audit of some processing effects in aggregated occurrence records. ZooKeys 751: 129-146. https://doi.org/10.3897/zookeys.751.24791
- Morris P, Lowery D, McPhillips T (2017) FilteredPush/FP-Akka: FP-Akka v1.6.1. Zenodo (2017-11-29). https://doi.org/10.5281/ZENODO.1068326
- Morris P, Lowery D (2018) Kurator.org/ffdq-api: Release Of Kurator's FFDQ-API library version 1.0.4. Zenodo https://doi.org/10.5281/ZENODO.891414
- Morris P, Hanken J, Lowery D, Ludäscher B, Macklin J, McPhillips T, Wieczorek J, Zhang Q (2018) Kurator: Tools for Improving Fitness for Use of Biodiversity Data. Biodiversity Information Science and Standards 2 https://doi.org/10.3897/biss.2.26539
- Morris P, Lowery D, Morris R, McPhillips T, Dou L, Song T (2019) FilteredPush/FP-KurationServices: FP-KurationServices release version 1.1.8. Zenodo https://doi.org/10.5281/zenodo.3533267
- Morris PJ, Hanken JA, Kelly M, Lowery DB, Ludäscher B, Macklin JA, McCallum C, Morris RA, Song T, Sweeney P (2014) Integrating High Throughput Digitization with Distributed Software: Supporting Data Flows in the New England Vascular Plant Network with FilteredPush Technologies. Society for the Preservation of Natural History Collections 29th Annual Meeting Programme & Abstracts. p. 34.
- Morris PJ (2019) FilteredPush/event_date_qc: Release version 2.0.2 of the event_date_qc library, implementation of all the TG2 Core Time Tests. Release date: 2019-11-07. Zenodo https://doi.org/10.5281/zenodo.596795
- Ogden CK, Richards IA (1923) The meaning of meaning: a study of the influence of language upon thought and of the science of symbolism. Eighth. Harcourt, Brace & World, Inc, New York, 296-336 pp.
- Peterson AT, Navarro-Sigüenza A, Benítez-Díaz H (1998) The need for continued scientific collecting; a geographic analysis of Mexican bird specimens. Ibis 140 (2): 288-294. https://doi.org/10.1111/j.1474-919x.1998.tb04391.x
- Ponder WF, Carter GA, Flemons P, Chapman RR (2001) Evaluation of Museum Collection Data for Use in Biodiversity Assessment. Conservation Biology 15 (3): 648-657. https://doi.org/10.1046/j.1523-1739.2001.015003648.x
- Rowe R (2005) Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. Journal of Biogeography 32 (11): 1883-1897. https://doi.org/doi: 10.1111/j.1365-2699.2005.01346.x
- Sanderson R, Ciccarese P, Young B (Eds) (2017) Web Annotation Data Model: W3C Recommendation. 23 February 2017. World Wide Web Consortium URL: https://www.w3.org/TR/annotation-model/
- Saraiva AM, Chapman A (2013) Biodiversity Data Quality (BDQ) Interest Group Charter. https://www.tdwg.org/community/bdq/. Accessed on: 2019-11-11.

- Sato K (2012) An Inside Look at Google BigQuery. https://cloud.google.com/files/BigQueryTechnicalWP.pdf. Accessed on: 2020-1-30.
- Soberon J, Llorente J, Benitez H (1996) An International View of National Biological Surveys. Annals of the Missouri Botanical Garden 83 (4): 562. https://doi.org/10.2307/2399997
- Sonatype, Inc. (2011) Maven Central. http://search.maven.org/. Accessed on: 2020-1-30.
- Stein BR, Wieczorek JR (2004) Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1 https://doi.org/10.17161/bi.v1i0.7
- Stockwell DB, Beach J, Stewart A, Vorontsov G, Vieglais D, Pereira RS (2006) The use of the GARP genetic algorithm and Internet grid computing in the Lifemapper world atlas of species biodiversity. Ecological Modelling 195: 139-145. https://doi.org/10.1016/j.ecolmodel.2005.11.016
- Strong DM, Lee YW, Wang RW (1997) Data quality in context. Communications of ACM 40 (5): 103-110. https://doi.org/10.1145/253769.253804
- TDWG (2007) Biodiversity Information Standards (TDWG). http://www.tdwg.org/. Accessed on: 2017-7-14.
- Thackway R, Cresswell I (1992) Environmental regionalisations of Australia - A user-oriented approach. Environmental Resources Information Network, Australian National Parks and Wildlife Service, Canberra https://doi.org/10.13140/RG.2.1.2491.3765
- Veiga AK (2016) A conceptual framework on biodiversity data quality. Tese (Doutorado) [Doctoral Thesis]. Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Computação e Sistemas Digitais, 156 pp.
- Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D, Robertson TJ (2017) A conceptual framework for quality assessment and management of biodiversity data. PLOS ONE 12 (6): e0178731. https://doi.org/10.1371/journal.pone.0178731
- Wieczorek J (2006) Darwin Core Task Group Charter. https://web.archive.org/web/20170712133632/http://www.tdwg.org/activities/darwincore/charter/. Accessed on: 2017-9-13.
- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLoS ONE 7 (1): e29715. https://doi.org/10.1371/journal.pone.0029715
- Wiley EO, Peterson AT (2004) Biodiversity and the Internet: Building and Using the Virtual World Museum. In: Scharl A (Ed.) Environmental Online Communication. Advanced Information and Knowledge Processing. Springer, London. https://doi.org/10.1007/978-1-4471-3798-6_11
- Wilkinson M, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J, da Silva Santos LB, Bourne P, Bouwman J, Brookes A, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo C, Finkers R, Gonzalez-Beltran A, Gray AG, Groth P, Goble C, Grethe J, Heringa J, 't Hoen PC, Hooft R, Kuhn T, Kok R, Kok J, Lusher S, Martone M, Mons A, Packer A, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S, Schultes E, Sengstag T, Slater T, Strawn G, Swertz M, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (1). https://doi.org/10.1038/sdata.2016.18

- Yesson C, Brewer P, Sutton T, Caithness N, Pahwa J, Burgess M, Gray WA, White R, Jones A, Bisby F, Culham A (2007) How Global Is the Global Biodiversity Information Facility? PLoS ONE 2 (11): e1124. https://doi.org/10.1371/journal.pone.0001124

Supplementary materials

Suppl. material 1: Vocabulary of Terms used for the TDWG Task Group on Data Quality Tests and Assertions doi

Authors: Arthur Chapman, John Wieczorek, Lee Belbin, Paul Morris, Allan Koch Veiga
Data type: Vocabulary
Brief description: Vocabulary of Terms used for the TDWG Task Group on Data Quality Tests and Assertions, plus key additional terms from the Use Case Study. Ther terms are consistent with the terms used in the Fitness for Use Framework (Veiga *et al.* 2017)
Download file (18.23 kb)

Suppl. material 2: Data Quality Use Case Study Results doi

Authors: Emily Rose Rees and Miles Nicholls
Data type: Study Results
Brief description: Use cases were collected using a number of methods to maximise responses. Lead authors of papers published using data accessed via the Atlas of Living Australia (ALA) were contacted and asked to contribute their research data use cases, and a number of papers describing fitness for use determination were sent to the ALA Data Quality group. Fitness for use and quality check information from these papers were extracted and transferred to the use case library. These are the results of those surveys.
Download file (325.64 kb)

Suppl. material 3: Counts of occurrence records in 2019-04-15 snapshot of GBIF for date-related validation tests doi

Authors: John Wieczorek
Data type: Summary Results
Brief description: Counts of occurrence records in 2019-04-15 snapshot of GBIF-mediated data that fit the three categories of expected responses for each of the event date-related validation tests.
Download file (22.50 kb)

Suppl. material 4: TG2 Test Descriptions doi

Authors: Lee Belbin, Arthur Chapman, John Wieczorek, Paula Zermoglio and Paul Morris
Data type: Specifications
Brief description: Description and specifications for the tests following the conventions of the Fitness For Use Framework. This supplement is a copy of https://github.com/tdwg/bdq/blob/master/tg2/core/TG2_tests.csv as of commit 941e774 2019-Aug-20.
Download file (100.74 kb)

Endnotes

*1 Where dwc: means the namespace http://rs.tdwg.org/dwc/terms/. See the discussion in the Darwin Core RDF Guide (Darwin Core and RDF/OWL Task Groups 2015)

*2 We refer to tests meaning the three layers that span the framework (Veiga 2016, Veiga et al. 2017) as Measures, Validations, and in the third layer, Amendments. Elements in this third layer, however, in the description of Data Quality needs, are termed Improvements, i.e., descriptions of means by which arbitrary data might be improved for some use. However, in the description of Data Quality reports that are termed Amendments, these are proposals for how specific data can be changed to improve its quality for some use. In developing the tests and assertions we tend to informally use Amendments as a synonym for both Improvements and Amendments.